Empirical Beliefs

J.W. Lloyd

School of Computing
College of Engineering and Computer Science
The Australian National University

Incomplete Draft – May 19, 2022

Preface

Empirical beliefs are beliefs that are acquired from observations by an agent situated in an environment. Agents use empirical beliefs to track reality. This manuscript is a draft of a mathematical theory of empirical beliefs. In particular, it examines in detail the structure of empirical beliefs, and how to acquire and utilize them.

The account here of empirical beliefs is probabilistic and modal. Probability theory is used to model uncertainty about beliefs and provides a form of 'degree of belief'. Modal operators provide doxastic and temporal aspects of beliefs. The main contributions are the introduction of the concept of a schema from which empirical beliefs are obtained, the ability to acquire from observations beliefs that are conditional distributions, the sophistication of the representation language for empirical beliefs, and the ability to reason about such beliefs.

The book could be of interest to researchers in computer science, engineering, logic, or philosophy. In computer science, artificial intelligence and machine learning researchers are concerned about the problem of acquiring and utilizing a sophisticated and detailed model of the environment and other agents. This model is expressed as a collection of empirical beliefs that the agent acquires from observations, and uses to act and communicate. Thus the theoretical results of this book are directly applicable in agent applications. In addition, AI safety research could benefit from the precision and expressiveness of the theoretical formalism. In engineering, researchers in signal processing and control theory study stochastic filtering. Here, acquisition of empirical beliefs is also via stochastic filtering; however, the setting here generalizes that setting because the concept of a state distribution in stochastic filtering is a special kind of empirical belief as defined here. In logic, researchers are concerned about the use of logic for knowledge representation. Here, reasoning is carried out in an unusually expressive logic, namely, modal higher-order logic, which admits the direct modelling of probabilistic, doxastic, and temporal aspects of empirical beliefs. Usually, modal logic is used to *analyze* agent systems; in contrast, here, modal higher-order logic is used as the language in which beliefs are represented. In philosophy, the ideas in the book could be useful to epistemologists in that they provide a precise definition of the concept of an empirical belief that has considerable generality and naturalness, and hence could be used to concretize epistemological theories. Also, the approach of stochastic filtering, used here to acquire empirical beliefs, takes a particular philosophical position on belief acquisition that would be interesting to investigate. Furthermore, the highly expressive logic in which beliefs are expressed provides opportunities for investigations in formal epistemology.

A primary goal of the theory is to use it in practical applications, and a number of novel results show promise for that. Potential applications include robotics, autonomous

vehicles, home automation, smart grids, and virtual personal assistants.

The book consists of five chapters and two appendices. The chapters contain the core material. To avoid interrupting the flow of the core material, two extensive appendices contain the necessary mathematical background on probability and logic to support the key results. The theoretical results are presented in a technically precise style. Wherever appropriate, examples and diagrams help provide the intuition behind the theoretical results.

The first chapter provides an overview of the contents of the book. The second chapter is concerned with state distributions, the prototypical kind of empirical belief. The third chapter studies the structure of empirical beliefs. The fourth chapter shows how to acquire empirical beliefs. The fifth chapter presents the logical representation of beliefs and shows how to reason with beliefs.

The first appendix gives background material on the relevant aspects of probability theory, especially probability kernels and regular conditional distributions. The second appendix gives the syntax and semantics of the logic, and describes how computation and proof are carried out. It also presents structural induction.

The two main concepts of the book are those of schema and empirical belief, where empirical beliefs are obtained from schemas by instantiating them with the current history. Schemas are sequences of regular conditional distributions, the definition of which provides a criterion for the correctness of schemas. Since regular conditional distributions are primarily dependent on the concept of conditional expectation, the latter concept appears prominently throughout the theory.

Here is a summary of the main contributions of the book. Overall, the book provides a mathematical theory of empirical beliefs. Its theme is doxastic rationality, that is, the ability to acquire beliefs that capture aspects of the environment as accurately as possible given the available observations. The definitions of schema and empirical belief are given that emphasize the important correctness property that the concept of a regular conditional distribution provides. The practical importance of dealing with probability measures over structured spaces in the codomains of empirical beliefs is explained. Based on these definitions and the need to handle structured spaces, a theoretical account of the construction and deconstruction of schemas and empirical beliefs is provided. The recurrence equations for stochastic filtering of schemas and empirical beliefs are established. Stochastic filtering is a natural method for acquiring empirical beliefs.

A suitable logic for logicizing empirical beliefs and reasoning about them is introduced. This logic is highly expressive and supports the reasoning needed for an agent to use empirical beliefs for the selection of actions. The basic theoretical results concerning the computation and proof aspects of the logic are established. Reasoning systems for the logic (in various forms) have had prototype implementations over the last 20 years. The evidence from these experiments suggests that the reasoning system presented here does seem to be feasible and practical.

In addition to a systematic theoretical account of empirical beliefs, contributions of note are the following:

1. The definitions of schema and empirical belief. (Definitions 3.1.1 and 3.1.3).

2. The results concerning the construction and deconstruction of schemas and empirical beliefs. (Propositions 3.2.1 to 3.3.5 and 3.3.1 to 3.3.4.)

3. The filter recurrence equations for schemas and empirical beliefs in the conditional case. (Propositions 4.2.3 and 4.2.11.)
4. The results showing that the environment can be synthesized from a schema and the transition and observation models for the schema in the conditional case. (Propositions 4.2.4 and 4.2.13.)
5. The observation model synthesis results. (Propositions 4.2.5 and 4.2.12.)
6. Bayesian inference is a special case of filtering. (Propositions 4.1.7 and 4.2.9.)
7. The algorithm for a conditional particle filter. (Figures 4.20 and 4.21.)
8. The algorithm for a factored conditional particle filter. (Figures 4.31 and 4.32.)
9. The definition of the denotation of a modal term. (Part 6 of Definition B.2.10.)
10. The definitions of computation of rank 0 (Definition B.3.1), proof of rank 0 (Definition B.3.4), computation of rank k (Definition B.3.6), and proof of rank k (Definition B.3.7).

Kee Siong Ng contributed significantly to this book through a series of papers that we wrote on the material of Chapter 5 and Appendix B.3. The results of a collaboration with Dawei Chen, Samuel Yang-Zhao, and Kee Siong Ng on a more extensive account of filtering algorithms than is currently presented in Chapter 4 and their application to modelling epidemic processes will appear elsewhere.

In its present form, this manuscript is a snapshot of an on-going research endeavour. For some sections there is still much work to be done. My intention is to post regular updates over the next couple of years. Comments, suggestions, and corrections are greatly appreciated. Finally, notwithstanding the earlier experimental work on reasoning in the logic and the more recent filtering experiments mentioned above, there remains much more experimental work that needs to be done to demonstrate that the machinery for acquiring and utilizing empirical beliefs proposed here pays sufficient dividends in terms of intelligent behaviour of agents. I would be pleased to hear from anyone interested in pursuing such experimental work.

Sydney, May 2022 John Lloyd

Contents

Chapter 1

Introduction

This chapter sets the scene with an informal discussion of the main ideas in the book. The first section makes some brief remarks about the philosophical underpinnings of the book. The second section describes the role of probability and logic in the theory of empirical beliefs. The next section describes the setting for agents and environments that will be employed. Empirical beliefs, which are beliefs having a particular form that an agent learns from observations, are then introduced. Subsequent sections discuss the acquisition and utilization of empirical beliefs.

1.1 Artificial Doxastic Rationality

Two components of rationality can be distinguished. *Doxastic rationality* concerns acquiring beliefs that capture aspects of the environment as accurately as possible given the available observations. ('Doxastic' derives from the Greek $\delta\acute{o}\xi a$, *doxa*, meaning 'opinion' or 'belief'.) *Instrumental rationality* concerns acting in such a way as to maximize performance measure. ('Instrumental' means 'serving as an instrument or means to achieve a particular end or purpose'.) Together these comprise rationality:

$$\text{Rationality} = \text{Doxastic Rationality} + \text{Instrumental Rationality}.$$

Instrumental rationality ensures an agent acts appropriately to achieve its goals based upon a model of the environment that doxastic rationality ensures is as accurate as the observations available to the agent allow.

Actually, the decomposition of rationality into doxastic rationality and instrumental rationality is not standard; the usual decomposition is into *epistemic* rationality and instrumental rationality. ('Epistemic' derives from the Greek $\grave{\epsilon}\pi\iota\sigma\tau\acute{\eta}\mu\eta$, *episteme*, meaning 'knowledge'.) Now epistemic rationality concerns knowledge while doxastic rationality concerns beliefs, so it is necessary to explain why the emphasis here is on beliefs rather than knowledge. Conventionally, knowledge is taken to be justified, true belief (leaving aside the debate about the Gettier cases). For rationality, justified belief is certainly necessary: the agent should have good reasons for its beliefs. And some beliefs will certainly be true. For example, function definitions, such as sorting and searching functions, in the code of an (artificial) agent, count as beliefs and are true. Also some knowledge of the world would be built into an agent by the designer. (Here, 'designer' means the person or

group of people who design and implement the agent.) But, for many beliefs, truth is too much to expect. For example, empirical beliefs, those that are learned from observations, generally will not be strictly true. This is because the observations available to learn the belief may be few and uncertain. In such cases, truth is replaced by degree of belief. The form of empirical beliefs adopted in this book explicitly indicates the degree of belief that should be ascribed to a belief and this degree is largely determined by the observations that are available to the agent: when the observations are sparse and less informative, the degree of belief will be low; when the observations are plentiful and more informative, the degree of belief will be high.

The general topic of this book is *artificial doxastic rationality*, that is, doxastic rationality exhibited by artificial agents. In particular, the book presents a theory of empirical beliefs that underlies doxastic rationality. Its main goals are to describe the structure of empirical beliefs, show how an agent can acquire empirical beliefs from observations, and explain how an agent can utilize empirical beliefs to act rationally.

1.2 Probability and Logic

The underlying technologies of the account here of empirical beliefs are probability theory and logic. The different roles played by each of these is now discussed.

Probability theory is a theory of uncertainty, and the most successful and widely used such theories. Here, probability theory is used in the standard way, although the heavy emphasis on probability kernels, needed for belief representation, is unusual. Logic provides the the machinery of reasoning, once again in a standard way, although the use of modal higher-order logic for belief representation is also unusual. However, there is a contrast in the way each of probability theory and logic are used in the theoretical and practical account of empirical beliefs that needs to be explained. This difference is intimately connected to the notions of semantics and syntax in logic.

In a typical application of logic, there is an intended interpretation. This can be thought of as a formalized description of the application as an interpretation of the appropriate type for the particular logic. The interpretation is called intended because it is a description of the actual application. An interpretation lives in the semantic component of the logic: it consists of the definitions of concrete domains and functions using ordinary mathematical concepts. Belief acquisition can be thought of as a way of constructing (part of) the intended interpretation: the definition of an empirical belief, a particular function in the intended interpretation, is acquired from observations. The belief acquisition method employed here is stochastic filtering, which includes Bayesian inference and therefore Bayesian machine learning methods as special cases. Other parts of the intended interpretation, including the definitions of non-empirical beliefs, are provided by the designer. Commonly, the intended interpretation is at best only informally stated. However, for the approach here, some formality is needed because a theory for which the intended interpretation is a model has to be constructed automatically from the intended interpretation in order to do reasoning.

On the other hand, the (logical) theory for an application lives in the syntactic component of the logic: a suitable theory consists of formulas that are valid in the intended interpretation. Proving that a formula is a theorem then shows that it is a logical conse-

quence of the theory, that is, valid for every interpretation that makes all the formulas in the theory valid. Since the intended interpretation has this property, it follows that the theorem is valid in the intended interpretation, a fact that can then be used to help select actions, for example. Thus reasoning is used to establish (as acccurate as possible) information about the actual application beyond that explicitly provided by beliefs: in order to act effectively, the agent needs to know as much as possible about its environment.

Now consider a typical artificial intelligence application. Often such applications, particularly machine learning ones, use the technology of probability, and make no mention of logic. In this case, the application can be considered as being entirely encompassed inside the ordinary mathematical setting of the intended interpretation, even if this is not formalized as a logical construct. Other typical artificial intelligence applications, particularly ones primarily concerned with knowledge representation, are presented completely in the setting of some logic, often first-order or modal propositional. In such cases, usually a suitable theory is presented and everything takes place inside the syntactic setting of the logic, with the intended interpretation only ever implicitly present.

In this book, the theoretical and practical development takes place in both the semantic and syntactic settings. The setting for Chapters 2, 3, and 4 is the semantic setting of ordinary mathematics, especially probability theory, while the setting for Chapter 5 is the syntactic setting of modal higher-order logic, a logic that is expressive enough to formalize probabilistic constructs by virtue of its higher-orderness. The transition from the domains and functions of the intended interpretation in the semantic setting to the corresponding formulas of the logical theory in the syntactic setting is intuitively obvious, but needs to be automated in an application to facilitate reasoning about empirical beliefs.

In summary, acquisition is for constructing (part of) the intended interpretation; reasoning is for discovering (implicit) information about the intended interpretation.

1.3 Agent-environment Systems

The setting for studying empirical beliefs is that of agent-environment systems that consist of an agent situated in some environment. From now on, 'agent' means an 'artificial agent'. An agent is a computer system that operates in some environment, receiving observations from the environment, and applying actions to the environment in order to achieve some goal(s). Typical agents are robots, autonomous vehicles, automated trading agents, and personal assistants. A key component of an agent that is used to select actions is its belief base which is the collection of all its beliefs. Particular attention will be paid here to empirical beliefs, the collection of which is called the empirical belief base. Figure 1.1 illustrates at a high level the interaction between an agent and its environment.

An agent-environment system can be modelled stochastically by a basic probability space $(\Omega, \mathfrak{S}, \mathsf{P})$. Intuitively, Ω is the set of all possible runs that the system may follow, where a run is the particular temporal sequence of events that occur. The probability measure P defined on the set $\mathfrak{S}$ of measurable subsets of Ω gives a distribution on all such runs. No further details about $(\Omega, \mathfrak{S}, \mathsf{P})$ are specified. Instead, all attention is concentrated on two stochastic processes:

$$\mathbf{a} : \Omega \to A^{\mathbb{N}}$$

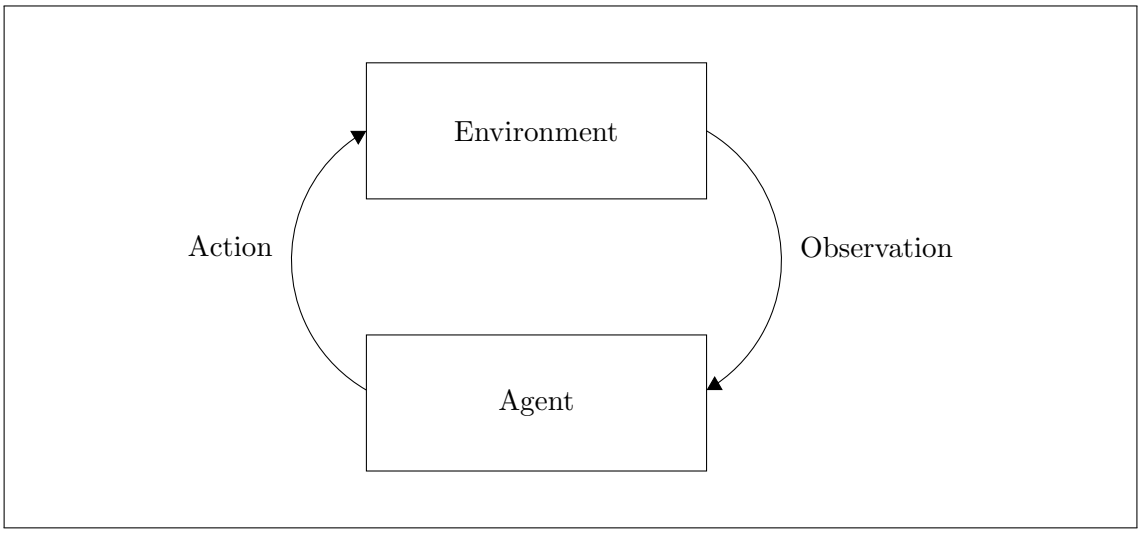

Figure 1.1: An agent and an environment

and

$$\mathbf{o} : \Omega \to O^{\mathbb{N}}.$$

Here, $\mathbb{N}$ is the set of positive integers, A is the space of actions, and O is the space of observations. Thus $A^{\mathbb{N}}$ can be regarded as the set of (infinite) sequences of elements in A; similarly, for $O^{\mathbb{N}}$. The stochastic process $\mathbf{a}$ can be decomposed into a sequence of random variables $\mathbf{a}_n : \Omega \to A$, for all $n \in \mathbb{N}$, such that $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \ldots)$; similarly, for $\mathbf{o}$. Thus, given a run $\omega \in \Omega$, $\mathbf{a}(\omega) = \mathbf{a}_1(\omega)\,\mathbf{a}_2(\omega)\,\mathbf{a}_3(\omega)\ldots$ is a sequence of actions. Similarly, $\mathbf{o}(\omega) = \mathbf{o}_1(\omega)\,\mathbf{o}_2(\omega)\,\mathbf{o}_3(\omega)\ldots$ is a sequence of observations. (See Figure 1.2.) The process $\mathbf{a}$ is called the *action process* and $\mathbf{o}$ is called the *observation process*.

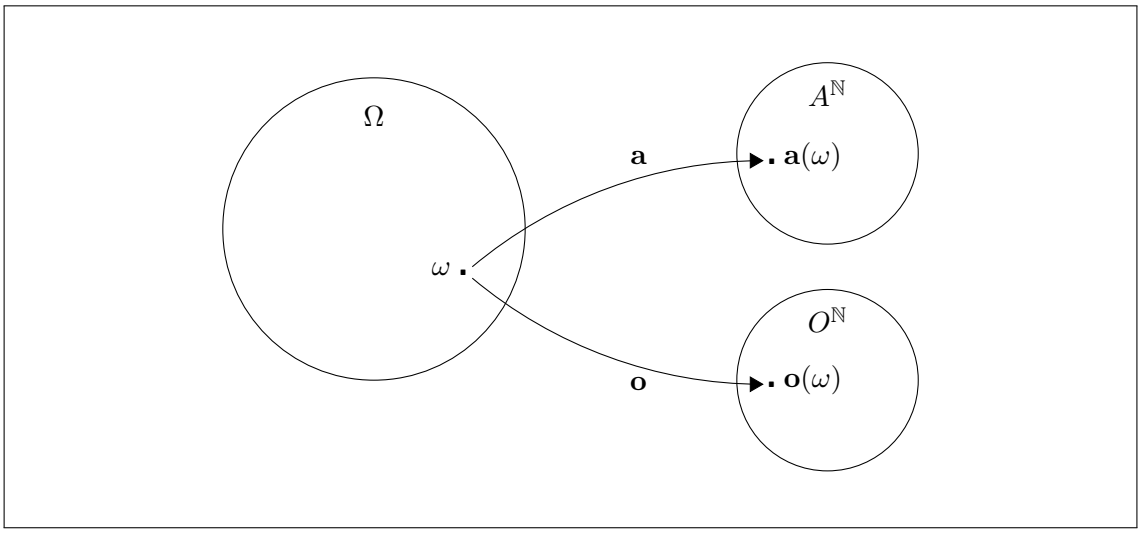

Figure 1.2: Basic probability space and the stochastic processes $\mathbf{a}$ and $\mathbf{o}$

Formally, the basic probability space $(\Omega, \mathfrak{S}, \mathsf{P})$, the spaces A and O, and the stochastic

processes **a** and **o** are collectively an *agent-environment system*. Usually the system will include other stochastic processes on other spaces as well.

For all $n \in \mathbb{N}_0$,

$$H_n \triangleq A \times O \times \cdots \times A \times O,$$

where there are n occurrences of A and n occurrences of O. (The symbol $\triangleq$ means 'stand(s) for'.) Here, $\mathbb{N}_0$ is the set of non-negative integers. Each H_n is the set of histories of action-observation cycles up until the end of the nth cycle, that is, time step n. All an external observer of the agent-environment system can observe is the history up to the current time step.

Considered abstractly, an agent is a sequence $\Lambda \triangleq (\Lambda_n)_{n\in\mathbb{N}}$ where

$$\Lambda_n : H_{n-1} \to \mathcal{P}(A)$$

is a (measurable) function, for all $n \in \mathbb{N}$. Here $\mathcal{P}(A)$ is the set of probability measures on the set of measurable subsets of A. However, Λ is not just any sequence of (measurable) functions of the form $\Lambda_n : H_{n-1} \to \mathcal{P}(A)$: each Λ_n must be 'consistent' with **a** and **o** in a way that is described in Section 2.1. An agent takes as input the history so far up to the last observation emitted by the environment and computes a distribution on the actions that could be applied by the agent. The agent then draws some action from this distribution. This abstract view of an agent models what would be seen by an external observer of an agent-environment system. In respect of the agent part of the system, given the history H_{n-1}, the observer sees only the action selected by the agent from the distribution $\mathcal{P}(A)$.

Considered abstractly, an environment is a sequence $\Xi \triangleq (\Xi_n)_{n\in\mathbb{N}}$, where

$$\Xi_n : H_{n-1} \times A \to \mathcal{P}(O)$$

is a (measurable) function, for all $n \in \mathbb{N}$. Here $\mathcal{P}(O)$ is the set of probability measures on the set of measurable subsets of O. Also, Ξ is not just any sequence of (measurable) functions of the form $\Xi_n : H_{n-1} \times A \to \mathcal{P}(O)$: each Ξ_n must be 'consistent' with **a** and **o** in a way that is described in Section 2.1. An environment takes as input the history so far and the last action applied by the agent and returns a distribution on the observations that could be emitted by the environment. The environment then draws some observation from this distribution. This abstract view of an environment models what would be observed by an external observer of an agent-environment system. In respect of the environment part of the system, given the history H_{n-1} and the action just selected by the agent, the observer sees only the observation selected by the environment from the distribution $\mathcal{P}(O)$.

The form taken by the observations in the space O depends on the application. For a mobile robot, they may simply be images taken by a camera. Probability measures over such low-level observations may be difficult to deal with, so it may be useful to map from low-level observations to higher-level observations with more structure. High-level observations for a mobile robot can be obtained by scene-recognition techniques from computer vision research. Such a high-level observation could consist of a set of objects, the spatial relationships between them, and so on. Observations in this form are highly suitable for the techniques proposed in this book. Intuitively, one could imagine such a high-level description of an observation being a suitable basis for a discussion between the

robot and a human about the observation; this is not the case for the observation as a set of pixels. Recently, there have been significant advances in the application of deep learning to scene-recognition tasks. So this provides one possible connection between the technology of deep learning and the technology proposed in this book: in effect, the classification function obtained by a deep-learning system for the scene-recognition problem becomes a (crucial) part of the stochastic process $\mathbf{o} : \Omega \to O^{\mathbb{N}}$, the part that gets from a set of pixels to a high-level description of the observation. In gaming-playing applications, such as for Poker or Go, observations come naturally in structured form that is directly useful for the techniques of this book. Similiarly, for an artificial personal assistant, observations are naturally structured.

In summary, what is ultimately provided to an observer of an agent and environment is an interleaved sequence

$$a_1\, o_1\, a_2\, o_2\, a_3\, o_3 \ldots$$

of actions $a_1, a_2, a_3, \ldots$ and observations $o_1, o_2, o_3, \ldots$ which come from some run $\omega \in \Omega$ that has been drawn from the distribution P and the stochastic processes $\mathbf{a}$ and $\mathbf{o}$.

1.4 Empirical Belief Structure

This section introduces empirical beliefs. But before that it is necessary to explain what a belief is.

A *belief* is a function that an agent uses to assist in choosing its actions. It is said that the agent *holds* the belief. Here, the meaning of 'function' is to be taken in the mathematical sense: a function $f : X \to Y$ is a subset $f \subseteq X \times Y$ such that, for each $x \in X$, there is one, and only one, $y \in Y$ satisfying $(x, y) \in f$. Usually, $(x, y) \in f$ is written $f(x) = y$. X is called the domain of the function, Y the codomain, and $X \to Y$ the signature of the function. The collection of beliefs held by an agent is called its *belief base*.

Some beliefs have a signature of the form

$$X \to \mathcal{P}(Y),$$

where X and Y are measurable spaces and $\mathcal{P}(Y)$ is the set of probability measures on Y. (A measurable function having signature of the form $X \to \mathcal{P}(Y)$ is called a probability kernel.) An important special case is when X is a distinguished singleton set. In this case, an empirical belief with signature $X \to \mathcal{P}(Y)$ can be identified with a distribution in $\mathcal{P}(Y)$, so that beliefs include the common case where a belief is a distribution on the state. If $\nu : X \to \mathcal{P}(Y)$ is a belief, then, for each $x \in X$, $\nu(x)$ is a probability measure on Y which thus provides a degree of belief about the value of the belief at x.

Commonly, a signature for a belief has one of the forms

$$X \to \mathcal{P}(\prod_{i \in I} Y_i),$$
$$X \to \mathcal{P}(\coprod_{i \in I} Y_i), \text{ or}$$
$$X \to \mathcal{P}(W^Z).$$

That is, the space supporting the probability measures in the codomain is either a product, a sum, or a function space, which is a special case of a product. In Chapter 3, it is shown how various data types of interest including tuples, lists, strings, sequences, sets, multisets, and graphs can be modelled using the mathematical concepts of products, sums, and function spaces.

The most natural, general, and direct way of defining probability distributions is via probability measures and hence these are used throughout the book, especially for the theoretical development. However, for practical applications, densities (that is, probability density functions) are often more convenient. (An exception is the application of particle filters in Chapter 4 for which probability measures in the form of mixtures of Dirac measures are appropriate.) Thus the theoretical development of densities is also presented and some of the examples use densities rather than probability measures. Using densities, under weak conditions, a probability kernel is represented by, and can be directly obtained from, a conditional density having signature $X \to \mathcal{D}(Y)$, where $\mathcal{D}(Y)$ is the set of densities on Y (with respect to some underlying measure on Y).

Beliefs can be characterized as either those that are built into the agent before deployment by the designer of the agent or those that are learned by the agent during deployment. The beliefs built in by the designer will mostly be knowledge, not just beliefs, since they will be true. For example, they could include all the function definitions in the Haskell Prelude, which will be needed in many computations.

The beliefs of most interest in this book are those learned during deployment. An empirical belief is a belief that is learned by an agent during deployment from observations. As an example, consider the following situation. The success (in the sense of achieving its goal(s)) of an agent situated in some environment depends in part upon the agent being able to capture quite subtle properties of the environment. Thus an agent would typically try to at least model particular aspects of its environment, if not the entire environment. So a possible empirical belief is the state distribution, where a state can be thought of as a model of those aspects of the environment and agent that could affect the future. A state typically summarizes relevant aspects of the environment rather than attempting to model the entire environment. If the state space is denoted by S, then the signature of the corresponding empirical beliefs is $\mathcal{P}(S)$. As another example, a poker agent might try to learn the conditions under which an opponent might fold a hand. Naturally, the corresponding empirical belief varies from opponent to opponent. In summary, as will be shown in Chapter 3, signatures of the form $X \to \mathcal{P}(Y)$ are sufficient to cover the signatures of empirical beliefs of interest.

However, on closer examination, it turns out that the concept of an empirical belief is not quite the fundamental concept here. Consider an empirical belief $\nu : X \to \mathcal{P}(Y)$. The reason that the empirical belief concept is not quite fundamental is because ν depends on the history up to the current time; thus empirical beliefs are contingent on the actual history that took place. In fact, there is a noncontingent concept, that of a schema, from which empirical beliefs can be obtained by substituting the actual history. A schema is a sequence $(\mu_n)_{n \in \mathbb{N}_0}$ of probability kernels, where each

$$\mu_n : H_n \times X \to \mathcal{P}(Y)$$

captures the conditional probabilistic nature of the distribution on Y given values in H_n and X. For an intuitive understanding of the conditional probabilistic nature of schemas,

consider Figure 1.3.

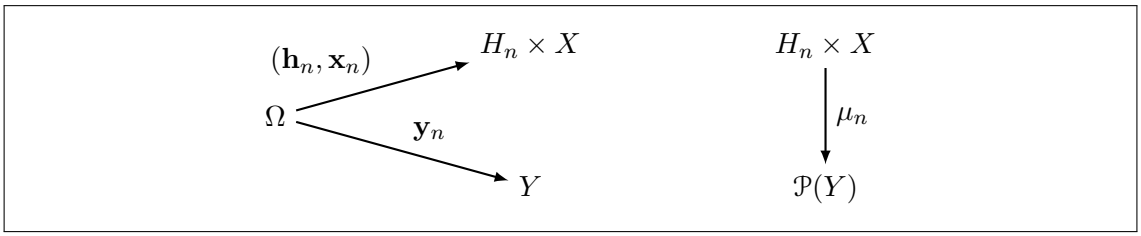

Figure 1.3: Setting for schemas

Here $(\Omega, \mathfrak{S}, \mathsf{P})$ is the basic probability space, A the action space, O the observation space, X and Y spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ the action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ the observation process, and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes. For all $n \in \mathbb{N}_0$, $\mathbf{h}_n \triangleq (\mathbf{a}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{o}_n)$. Note that, for all $\omega \in \Omega$, $\mathbf{h}_n(\omega) \in H_n$ is the history up to the end of n action-observation cycles.

A *schema* (*for* $\mathbf{y}$ *given* $\mathbf{x}$) is a sequence $\mu \triangleq (\mu_n)_{n \in \mathbb{N}_0}$, where

$$\mu_n : H_n \times X \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{x}_n)$, for all $n \in \mathbb{N}_0$. The (slightly imprecise) meaning of μ_n being a 'regular conditional distribution' is that

$$\mathsf{P}(\mathbf{y}_n^{-1}(B) \mid (\mathbf{h}_n, \mathbf{x}_n)) = \lambda\omega.\mu_n((\mathbf{h}_n, \mathbf{x}_n)(\omega))(B),$$

for all (measurable) subsets $B \subseteq Y$. The issue now is what does this equation mean? On the left-hand side is the conditional probability $\mathsf{P}(\mathbf{y}_n^{-1}(B) \mid (\mathbf{h}_n, \mathbf{x}_n))$. Intuitively, this is the conditional probability of $\mathbf{y}_n^{-1}(B) \subseteq \Omega$ given the information available from $(\mathbf{h}_n, \mathbf{x}_n)$. The equation shows that, for each B, the expression involving the probability kernel μ_n on the right-hand side gives the conditional probability on the left-hand side. So μ_n is not some arbitrary probability kernel; instead it captures a fundamental probabilistic property of the underlying agent-environment stochastic system. In fact, because of this property, μ_n is (essentially) unique.

This intuitive explanation of a schema glides over several technical issues. Conditional probabilities are special cases of conditional expectations, which are random variables; note that the expression on the right-hand side of the equation is indeed a random variable. Conditional expectation may not be an easy concept to understand but it is a fundamental concept in modern probability theory. It is just as crucial in the theory of empirical beliefs that follows: nearly every important result uses the fact that particular probability kernels are regular conditional distributions. The precise definition of a schema is given in Definition 3.1.1 and a detailed exposition of the requisite probability theory for this definition is given in Appendix A.

A schema is a noncontingent property of an agent-environment setting: for any history and any value in X, a schema determines the distribution on the possible values in Y. A typical example of a schema is a state schema $(\mu_n : H_n \to \mathcal{P}(S))_{n \in \mathbb{N}_0}$, where S is the set of states.

Empirical beliefs are obtained from schemas. An *empirical belief* is a probability kernel of the form

$$\lambda x.\mu_n(h, x) : X \to \mathcal{P}(Y),$$

where $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ is a schema and $h \in H_n$, for some $n \in \mathbb{N}_0$. (The precise definition of an empirical belief is given in Definition 3.1.3.) Thus an empirical belief with signature $X \to \mathcal{P}(Y)$ can be obtained from a schema by substituting the actual history observed into the first argument of the (corresponding component of the) schema. For example, from a state schema $(\mu_n : H_n \to \mathcal{P}(S))_{n\in\mathbb{N}_0}$, one can obtain an empirical belief, which is a state distribution, $\mu_n(h)$, where $h \in H_n$, for some $n \in \mathbb{N}_0$.

An empirical belief is a contingent property of an agent-environment setting: given the specific history that has taken place, for any value in X, an empirical belief determines the distribution on the possible values in Y. Thus empirical beliefs are contingent properties because they depend on the actual history observed. Note that empirical beliefs are justified (in an epistemological sense) since the actual current history is used to specialize a schema that correctly models the relevant aspect of the stochastic process determined by the agent and environment. Thus the justification is based on empiricism.

Note carefully that it is not the fact that the codomain has the form $\mathcal{P}(Y)$ that alone makes a belief empirical, but the fact that the belief depends upon the history. For example, the designer may build a belief having signature $X \to \mathcal{P}(Y)$ into an agent. From the agent's perspective, this belief is not empirical because it is not determined by observations. On the other hand, the agent could learn the belief during deployment in which case it would be an empirical belief.

In the literature, beliefs learned from observations often have a signature of the form $f : X \to Y$, where Y is not a space of probability measures. For example, most machine learning algorithms produce hypotheses in this form. However, it may be advantageous to learn an empirical belief with signature of the form $f : X \to \mathcal{P}(Y)$ even when the 'real' belief has signature $X \to Y$ because then the codomain of the empirical belief provides a degree of belief $f(x)$ for the possible values of the 'real' belief, for each $x \in X$. In fact, the definition of empirical beliefs above includes the case of those having signature of the form $X \to Y$. Consider the case where the value of an empirical belief for each $x \in X$ is a Dirac measure. Then such a belief having signature $X \to \mathcal{P}(Y)$ can be identified with one having signature $X \to Y$.

The understanding is that beliefs which are not empirical are actually knowledge and thus are true (and justified). Such beliefs are built into the agent before deployment by the designer. The responsibility for ensuring these beliefs are true falls to the designer. On the other hand, empirical beliefs are generally only approximations to the truth whose accuracy depends mainly on the precision and volume of observations.

Now consider an empirical belief $\nu : X \to \mathcal{P}(Y)$. In general, both X and Y can be structured. For example, X or Y could be a product of factors that could be lists or sets, each of which could have further structure inside. Consider first the case of X having structure and ask what form the function ν might take? In general, there are few possibilities for this. A useful form is that of a piecewise-constant function, which means that there exists a partition $\{X_i\}_{i=1}^n$ of X such that ν is constant (that is, takes a particular probability measure as a value) on each equivalence class of the partition. This means that when acquiring an empirical belief, one key issue is to find the 'right'

partition of X. The method adopted here is for the designer to specify a grammar for predicate construction which the learning system can employ. For each equivalence class, the agent searches systematically through the list of predicates given by the grammar and uses a heuristic to pick one that will be used to give the equivalence class. Note that the specification of the predicate grammar is an important method by which the designer provides knowledge of the environment to the agent.

Similarly, the space Y in the signature can have some structure. In this case, one approach is to deconstruct the empirical belief into a number of 'simpler' constituents. For example, a probability kernel $\nu : X \to \mathcal{P}(Y_1 \times Y_2 \times Y_3)$ can be deconstructed into

$$\nu_1 : X \to \mathcal{P}(Y_1)$$
$$\nu_2 : X \times Y_1 \to \mathcal{P}(Y_2)$$
$$\nu_3 : X \times Y_1 \times Y_2 \to \mathcal{P}(Y_3),$$

where ν is the product of ν_1, ν_2, and ν_3. If Y is a product space with high dimension, as can typically happen in applications, the empirical belief may have hundreds of factors, but each factor will have a codomain for which the support of the probability measures is 'simpler'. This is a trade-off that may be well worth making. Also, since $\mathbb{B}^Z$ can be regarded as the set of subsets of some set Z, where $\mathbb{B}$ is the booleans, this decomposition of product spaces is also relevant for the case of *countably infinite* products. However, as shall be seen, the motivation is not to consider infinite subsets of Z, but finite subsets of unbounded size. Another common case is when Y is a sum space (that is, a disjoint union of spaces). In this case, the empirical belief can deconstructed into a sum of 'simpler' constituents.

In summary, after decomposition, a typical empirical belief has a signature of the form $X \to \mathcal{P}(Y)$, where X structured but Y is not. The empirical belief may be piecewise-constant. Also, typically, Y is $\mathbb{R}$ (the set of real numbers) or $\mathbb{R}^k$, for some k, and the distributions of interest are Gaussian distributions on $\mathbb{R}$ or $\mathbb{R}^k$, Y is $\mathbb{N}_0$ and the distributions of interest are Poisson distributions on $\mathbb{N}_0$, or Y is a finite space and the distributions of interest are categorical distributions on Y. All these cases are about as simple as an empirical belief can be, but there may be hundreds of such beliefs that have to be processed in each time step of the agent program.

To provide an easier introduction to schemas and empirical beliefs, Chapter 2 covers the simpler nonconditional case of empirical beliefs that are state distributions. Then Chapter 3 covers the general case of schemas and empirical beliefs that are conditional.

1.5 Empirical Belief Acquisition

Belief bases are generally dynamic, that is, they change from time to time during deployment of the agent. It follows that agents need to have some method by which they can acquire beliefs during deployment. The phrase 'belief acquisition' is used to name this process. The term 'acquire' is intended to be understood in a general sense that includes 'filter', 'track', 'maintain', 'update', 'revise' and 'learn' as special cases. The term 'filter' refers to stochastic filtering, 'track' refers to the original application of filtering to tracking objects in flight, 'maintain' refers to the general process of maintaining a belief base over the life of the agent, 'update' and 'revise' refer to the forms of acquisition that

are studied in the literature on belief revision, and 'learning' refers to machine learning. The meaning of 'acquire' also includes the case when an agent decides during deployment that it needs a *new* kind of empirical belief (maybe not even anticipated by the designer) and goes about what is needed to learn it, altogether a form of invention. Thus the term 'acquisition' is intended to cover the spectrum of possible meanings of this term, including naive updating, the removal of inconsistency, the generalization that is characteristic of learning, and the invention of new concepts, and all of this over the lifetime of the agent. For empirical beliefs of the specific form considered here, the relevant approach to belief acquisition comes from signal processing research and is called stochastic filtering. In the following, stochastic filtering is generally referred to more simply as filtering.

Before discussing filtering in more detail, it is necessary to explain why this is essentially the only acquisition technique studied in this book, especially since the collection of methods commonly used for machine learning applications is now truly vast. Towards this, note that the only new information available to an agent during deployment comes from observations from the environment. Depending on the application, it may be possible, for example, for the agent to extract an i.i.d. data set from the observations to learn some function that would be useful to the agent. For such a task, any of the complete array of machine learning methods for i.i.d. data could be made available to an agent. Alternatively, an online learning method (different from filtering) could be used to learn a function that may change over time. All of this is taken for granted. This book instead concentrates on acquiring beliefs through a generalized filtering algorithm. This is partly because the filtering method has some novel aspects that I argue in this book are important for acquiring a good model of the environment, but more importantly because the empirical beliefs acquired this way seem essential to maintaining any kind of sensible behaviour of an agent. For example, imagine what would happen to an autonomous vehicle or a robot if its basic sense/think/act cycle was interrupted because the filter algorithm stopped and its empirical beliefs were no longer being updated. In other words, filtering seems essential to the continuing effective operation of most agents.

Consider the case when a schema has a signature of the form $H_n \to \mathcal{P}(Y)$, so that the corresponding empirical beliefs have a signature of the form $\mathcal{P}(Y)$. Thus this includes the case of an empirical belief that is a state distribution. Filtering is the standard method of updating a state distribution given the most recent action and observation. Under suitable conditional independence assumptions, this requires a transition model which is a function having a signature of the form $A \times Y \to \mathcal{P}(Y)$ that shows the effect of applying an action in A on a value in Y. It also requires an observation model which is a function having a signature of the form $Y \to \mathcal{P}(O)$. This shows what observation in O is likely to be observed for any given value in Y. Applying the current action via the transition model is called the transition update and taking into account the current observation via the observation model is called the observation update. Filtering, which consists of a transition update followed by an observation update, then provides an explicit expression for calculating the updated schema. From this, the updated empirical belief can be obtained by substituting the history up to the latest action and observation into the first argument of the updated schema.

In the more general case, when a component of a schema has a signature of the form $H_n \times X \to \mathcal{P}(Y)$, the filtering process can be extended. Both the transition model and observation model now depend on X and, in general, both depend on the history as well.

Here are the details.

A transition model is a sequence $(\tau_n)_{n\in\mathbb{N}}$ of probability kernels, where each τ_n is called a component of the transition model and has a signature

$$\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y).$$

A component of the transition model takes as input a value in $H_{n-1} \times A \times X \times Y$ and returns a distribution on the values in Y that could result from the transition. The A and the Y in the domain are standard; the X is in the domain because this is the transition model for the conditional case and the H_{n-1} is there because this covers cases more general than when Y is a state.

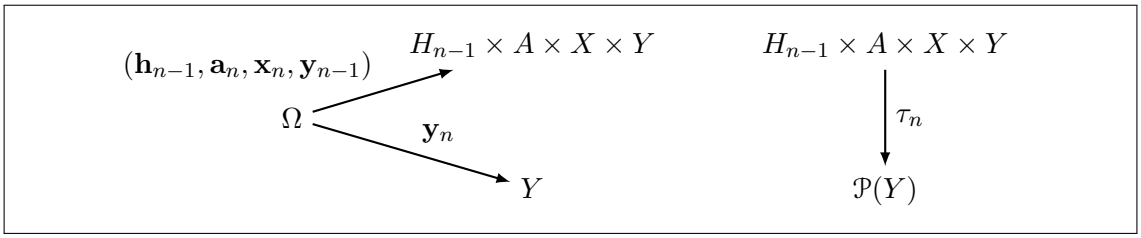

Figure 1.4: Setting for transition models

Now consider Figure 1.4. A *transition model* (*for* $\mathbf{y}$ *given* $\mathbf{x}$) is a sequence $\tau \triangleq (\tau_n)_{n\in\mathbb{N}}$, where

$$\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})$, for all $n \in \mathbb{N}$. The (slightly imprecise) meaning of this is that

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})) = \lambda\omega.\tau_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})(\omega))(C),$$

for all (measurable) subsets $C \subseteq Y$. Each τ_n is (essentially) unique.

An observation model is a sequence $\xi \triangleq (\xi_n)_{n\in\mathbb{N}}$ of probability kernels, where each ξ_n has signature

$$\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O).$$

An observation model takes as input a value in $H_{n-1}\times A\times X\times Y$ and returns a distribution on the observations that could be received by the agent. The Y in the domain is standard; the X is in the domain because this is the observation model for the conditional case and the $H_{n-1} \times A$ is there because this covers cases more general than when Y is a state.

Now consider Figure 1.5. An *observation model* (*for* $\mathbf{y}$ *given* $\mathbf{x}$) is a sequence $\xi \triangleq (\xi_n)_{n\in\mathbb{N}}$, where

$$\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O)$$

is a regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)$, for all $n \in \mathbb{N}$. The (slightly imprecise) meaning of this is that

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)) = \lambda\omega.\xi_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)(\omega))(B),$$

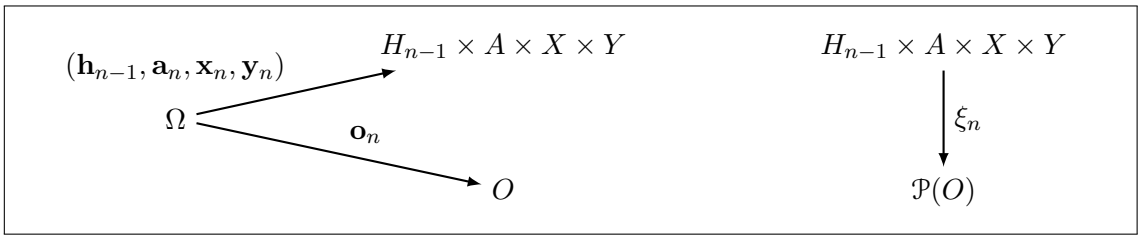

Figure 1.5: Setting for observation models

for all (measurable) subsets $B \subseteq O$. Each ξ_n is (essentially) unique.

Then, according to the filter recurrence equations of Chapter 4, the schema component $\mu_n : H_n \times X \to \mathcal{P}(Y)$ at time n is updated to the schema component $\mu_{n+1} : H_{n+1} \times X \to \mathcal{P}(Y)$ at time $n+1$. Correspondingly, the empirical belief $\lambda x.\mu_n(h_n, x) : X \to \mathcal{P}(Y)$ at time n is updated to the empirical belief $\lambda x.\mu_{n+1}(h_{n+1}, x) : X \to \mathcal{P}(Y)$ at time $n+1$. There are various complications in this more general case, but the same essential idea of filtering state distributions still works.

Given the transition and observations models, and the initial distribution on Y, the belief acquisition process can be automated. But where do the transition and observation models can from? At present they need to be provided by the designer. Depending on the application, this may be an onerous task. To begin with much depends on the designer giving the *correct* transition and observation models for the application; that is, the regular conditional distribution conditions in the definitions of these models must be satisfied. Also it is envisioned that an agent may have tens or hundreds of different kinds of empirical beliefs, not just a single state distribution, and the transition and observation models for each of these kinds of empirical beliefs need to be given by the designer. This can be a major engineering task. And then there is the issue, not considered in detail here, of using the empirical beliefs to decide how to act that is another significant engineering task. However, note that for parametrized transition and observation models, it is possible to estimate the parameters, a topic discussed in Chapter 4; thus it only becomes necessary for the designer to give the general form of the transition and observation models. It may also be possible in future to learn the (general form of the) transition and observation models for some applications.

From a theoretical point of view, this provides a straightforward view of the main idea of filtering. However, in practice, there is a major problem: except in a few special cases, the expression for the updated empirical belief is not tractable. This means that, while there is an explicit mathematical expression for the updated empirical belief, it is not easily calculable. The problem is that, except in a few cases, the result of the transition update and the result of the observation update cannot be simplified, so that the syntactic size of the expression for the updated empirical belief is linear in the number of time steps. After even a small number of time steps, the expression becomes intractable. There are two important cases where simplification is possible and the resulting expression is tractable: one is where the state distribution is Gaussian and the transaction and observation models are linear Gaussian (the case of linear dynamical systems) and the other is where the state space is finite and the state distribution is categorical (the case of hidden Markov models). (Throughout, 'hidden Markov model' will mean 'finite-state-space hidden Markov model'.)

More generally, the expression for the updated empirical belief is usually not tractable and a different approach is necessary. The standard solution for state distributions is to use a particle filter. This approach has been so successful that particle filters are used almost universally in robotics and autonomous vehicle applications, and hence it is also explored in detail in this book. However, the use of particle filters means the exact distributions in the codomains of empirical beliefs are approximated by mixtures of Dirac measures. For computing integrals, the approximation is just as useful as the exact distribution. But, for other reasoning tasks, the use of a mixture of Dirac measures may limit what can be done.

Stochastic filtering includes Bayesian inference as a special case. Basically, for Bayesian inference, the transition model is not needed and the stochastic processes $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ have the property that they are (roughly speaking) constant. This means (roughly) that $\mathbf{x}_n = \mathbf{x}_{n+1}$, for all $n \in \mathbb{N}_0$, and similarly for $\mathbf{y}$. Since Bayesian inference is a special case, this means all the methods of Bayesian machine learning, including Bayesian deep learning, can be brought to bear to acquire empirical beliefs. There are Bayesian versions of almost all machine learning methods. Thus essentially the whole gamut of machine learning methods are available for acquiring empirical beliefs. However, while Bayesian inference is simple in principle, when it comes to practice, approximations normally have to be made and then the situation can become rather complicated. For an illustration of this, see the chapter on Bayesian neural networks in [114, Ch.16].

To provide an easier introduction to the acquisition of empirical beliefs, Chapter 2 covers the simpler case of the acquisition of empirical beliefs that are state distributions. Then Chapter 4 covers the general case of the acquisition of conditional empirical beliefs.

1.6 Empirical Belief Utilization

Agents need beliefs to assist them in making effective choices of actions, which include communicating with other agents and humans. An agent's utilization of its beliefs requires two forms of reasoning: computation and proof. To reason, it is necessary for the beliefs to be represented in a formal language and for the reasoning process to be precisely formulated. In other words, a logic is needed so that beliefs can be represented and reasoning about beliefs can be carried out in the logic. So, at this point, the development passes from the semantic setting of interpretations of Sections 1.4 and 1.5 to the syntactic setting of theories in this section. The process of representation in a logic is called *logicization*. A suitable logic is modal higher-order logic, where modalities capture temporal, doxastic, and epistemic notions, and higher-orderness means that functions can take other functions as arguments.

One way to think about the higher-order part of the logic, also known as simple type theory, is that it is a formalization of everyday informal mathematics. Mathematical concepts are easy to express directly in higher-order logic because, amongst other things, the logic allows quantification over predicates and functions. This is illustrated in the agent programming context by the heavy and often essential use of higher-order functions in examples given in this book. In contrast, first-order logic only allows one to model many mathematical concepts indirectly and requires the introduction of (semantically complicated) set theory to give a satisfactory foundation for mathematics. The great ex-

pressive power of higher-order logic partly explains its widespread use in several subfields of computer science; in functional programming, where a program can be understood as a higher-order equational theory; in formal methods, where the logic is used to give specifications of programs and prove properties about them; in theoretical computer science, where various kinds of semantics are typically higher order; and elsewhere.

It is common for beliefs to have a modal nature, usually temporal or doxastic. For example, on the temporal side, it might be important that at the last time step or at some time in the past, some situation held and, therefore, a certain action is now appropriate. Similarly, on the doxastic side, beliefs about the beliefs of other agents may be used to determine which action to perform. The usefulness of modal beliefs for agents is now well established. Besides, introspection reveals that people use temporal and doxastic considerations when deciding what to do. A major trend in agent research is the use of modal logic to analyse properties of agents. In contrast, this book aims to use modal logic to *build* agents rather than analyse them. An important consequence of this objective is that the modal *propositional* logic that is generally sufficient to analyse agents is replaced here by modal *higher-order* logic that is needed to represent beliefs of agents performing non-trivial tasks.

While modalities can have a variety of meanings, what are of most concern here are doxastic, epistemic, and temporal modalities. For each agent i in a system, there is a corresponding doxastic or epistemic operator, typically either $\boldsymbol{B}_i$ or $\boldsymbol{K}_i$. If φ is a formula, then the informal meaning of $\boldsymbol{B}_i\varphi$ is 'agent i believes φ', while the informal meaning of $\boldsymbol{K}_i\varphi$ is 'agent i knows φ'. These modalities allow one to capture quite subtle facts about the beliefs and knowledge of agents. For example, the meaning of

$$\boldsymbol{B}_i\varphi \longrightarrow \boldsymbol{B}_i\boldsymbol{B}_j\varphi$$

is that 'if agent i believes φ, then agent i believes that agent j believes φ'. Temporal modalities are also useful. For example, here are some typical (past) temporal modalities: $\bullet$ ('last'), $\blacksquare$ ('always in the past'), $\blacklozenge$ ('sometime in the past'), and $\boldsymbol{S}$ ('since'). (There are also future versions of these.) Thus the informal meaning of $\bullet\varphi$ is that 'φ held at the last (that is, immediately preceding) time'. Putting the two kinds of modalities together, there are formulas such as

$$\bullet\boldsymbol{B}_i\varphi \longrightarrow \boldsymbol{B}_i\bullet\varphi$$

whose informal meaning is 'if, at the last time, agent i believed φ, then agent i believes (now) that φ held at the last time'. As part of an agent's belief theory, this kind of belief formula is genuinely useful for determining which action to perform next, as is shown later.

In most agent applications, it is necessary for the agent to deal with uncertainty. This issue leads directly to the more general problem of integrating logic and probability, a topic in artificial intelligence that is currently attracting substantial interest. One of the advantages of working in a higher-order logic is that it is expressive enough to easily encompass uncertainty without any additional logical machinery. In particular, the higher-orderness of the logic allows the representation of empirical beliefs having a signature of the form $X \to \mathcal{D}(Y)$ directly. While higher-orderness allows the representation of probabilities, extra reasoning capabilities beyond those normally provided by higher-order functional programming languages are needed to make the logic truly probabilistic.

Here are some remarks to clarify the distinction between computation and proof. First, consider computation. Intuitively, computation is the process of reducing an expression to a form which cannot be reduced any further and hence can be considered to be the value of the original expression. As an example, consider the expression $append([1,2,3],[4,5])$, where *append* is the function that appends two lists together. The expression can be reduced by a sequence of computation steps to the expression $[1,2,3,4,5]$, which can then be regarded as the value of the original expression. Computation is what functional programming languages such as Haskell do. Another example relevant to agent applications is that of computing the expected utility of a state. This computation reduces an integral to a value that is the expected utility.

Proof just means the standard concept of showing some formula can be proved from some set of assumptions. The proof system employed by the logic introduced here is a tableau system. For example, under the assumptions $\forall x.(p(x) \longrightarrow q(x))$ and $\forall x.(q(x) \longrightarrow r(x))$, there is a proof of $\forall x.(p(x) \longrightarrow r(x))$. More generally, computation and proof can be combined, so a proof can invoke a computation at some point and a computation can invoke a proof at some point. For agent applications, computations are most common, with subsidiary proofs sometimes needed in the computation.

Agents need to be able to build mathematical models in order to understand the world, to explain what they perceive, to imagine what might happen next, and to select actions to achieve goals. As explained in Section 1.2, representing and acquiring empirical beliefs take place in the semantics of the logic; this is literally building interpretations in a suitable logic. The language of interpretations is essentially the ordinary language of mathematics. The utilization of empirical beliefs then involves reasoning in a theory in the logic for which the intended interpretation is a model. This means that each formula in the theory is valid in the intended interpretation. Then a soundness result shows that every formula obtained by reasoning using this theory is valid in the intended interpretation. Thus the results of reasoning are correct. The construction of this theory from the intended interpretation must be automatable.

For the kind of applications envisioned here, this theory must be a theory in a highly expressive logic. Modal higher-order logic is an excellent candidate for such a logic. This logic subsumes first-order logic, the standard logic for knowledge representation in artificial intelligence; subsumes modal propositional logic, the standard logic for analysing multi-agent and other dynamic systems; allows the direct formalization of probabilistic concepts through its higher-orderness; and, with additional formalism, can encompass specialized knowledge representation languages such as the situation calculus, description logics, and causal logics.

In summary, the focus here is on representing and acquiring empirical beliefs in the language of everyday mathematics and then utilizing those empirical beliefs by reasoning in the associated theories in modal higher-order logic.

Bibliographical Notes

These notes begin with a general discussion about rationality and its relationship to intelligence.

The normative concept of rationality is used in psychology, economics, artificial intel-

ligence, and elsewhere. Roughly speaking, an (artificial or human) agent is rational if it chooses actions that maximize its performance measure. Since rationality is a normative concept, it sets a standard that agents *should* achieve. Here is a somewhat more precise definition, taken from [140, p.37]:

> "For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has".

The usual definition of performance measure is expected utility, where utility is a measure of the value of a state of the environment that an agent may find itself in after performing some action [140, Ch. 16]. The use of a utility function is well motivated by the Morgenstern-von Neumann theorem [161]. For this, the concept of a lottery, which summarizes the outcome of an action, is introduced. A lottery with possible outcomes (that is, states that could occur as a result of the action) $S_1, \ldots, S_n$ having probabilities of occuring $p_1, \ldots, p_n$, respectively, is denoted $[p_1, S_1; \ldots; p_n, S_n]$. Then a preference relation on lotteries can be introduced. The preference relation satisfies some so-called choice axioms that describe natural and intuitive properties that a putative preference relation should satisfy. The von Neumann-Morgenstern theorem states that under these conditions there exists a real-valued utility function U on the set of lotteries (and hence the states) that is consistent with the preference relation in the sense that, for all lotteries L and M, $U(L) > U(M)$ if and only if L is preferred to M. Furthermore, $U([p_1, S_1; \ldots; p_n, S_n]) = \sum_{i=1}^{n} p_i U(S_i)$, for all lotteries. Thus the use of utility functions instead of preference relations over the lotteries is justified.

One can even give a precise definition of a measure of how rational an agent is. In [92], the (measure of) rationality $\Upsilon(\pi)$ of an agent π is defined to be

$$\Upsilon(\pi) = \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi.$$

Here E is the set of all environments, $K(\mu)$ is the Kolmogorov complexity of μ, and V_μ^π is the expected cumulative reward that agent π can get in environment μ. The Kolmogorov complexity of a binary string x is defined as the length of the shortest program that computes x. Thus

$$K(x) = \min_p \{l(p) : \mathcal{U}(p) = x\},$$

where p is a binary string which is called a program, $l(p)$ is the length of this string, and $\mathcal{U}$ is a universal Turing machine. Intuitively, if a short program can be used to describe an environment, then the environment has low complexity. The term $2^{-K(\mu)}$ can hence be interpreted as the probability that the environment is μ. Thus $\Upsilon(\pi)$ can be understood as the expected cumulative reward for agent π, where the expectation is taken over all environments. (A number of technical details in this description have been elided; these are discussed in [92].) This is an elegant definition of the concept of rationality, but note that the concept is incomputable since Kolmorogov complexity is incomputable. However, the definition provides a gold standard against which more practical, computable definitions

of rationality can be compared. (By the way, in [92], $\Upsilon(\pi)$ is actually called the universal intelligence of agent π, not its rationality. More on this below.)

The concept of rationality is crucial in psychology, economics, artificial intelligence, and elsewhere. For example, the most influential textbook on artificial intelligence [140] essentially defines the field of artificial intelligence to be that of building rational agents. And much of microeconomics is concerned with understanding how rational agents make decisions in the market place. However, in psychology and elsewhere, the correctness of the assumption that people are rational has been seriously questioned. The research program investigating this issue is called the heuristics and biases research program, and was initiated by [157]. An accesssible account of the many experiments carried out that demonstrate this irrationality (that is, departure from the normative concept of rationality) can be found in [82]. As summarized in [151],

> " ... people assess probabilities incorrectly, they test hypotheses inefficiently, they violate the axioms of utility theory, they do not properly calibrate degrees of belief, their choices are affected by irrelevant context, they ignore the alternative hypothesis when evaluating data, and they display numerous other information processing biases".

There is continuing debate [155] about what level of rationality should be ascribed to the human mind. While it is hard to refute the empirical results of experimental psychologists that exhibit clear irrational behaviour of people, critics have found numerous alternative interpretations of the findings of the heuristics and biases research program that put human rationality in a better light. For example, some critics have suggested that rationality as defined above is too demanding as a normative requirement and should be weakened to something closer to a descriptive concept of actual human behaviour [150, 155].

Now the discussion turns to the closely related concept of intelligence, which intuitively means cognitive ability, but is notoriously difficult to define precisely. For example, [91] lists 70 definitions of intelligence that appear in dictionaries, encyclopedias, the psychological literature, and the artificial intelligence literature. These definitions vary greatly; here are three typical ones.

> "Intelligence is what is measured by intelligence tests" [20].
>
> "Intelligence is a very general mental capability that, amongst other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" [63].
>
> "Intelligence measures an agent's ability to achieve goals in a wide range of environments" [91].

What is the relationship between intelligence and rationality? In folk psychology, intelligence and rationality are distinguished. Intuitively, intelligence is understood to mean raw intellectual ability while rationality means doing the right thing or at least not doing stupid things. Psychologists make this distinction as well, but in a much more nuanced way of course. For discussions of the relationship between rationality and intelligence, see [150], which contains a comprehensive list of relevant papers, and [151]. A concise but useful summary of the main points can be found in [68].

While there are reasonably precise definitions of various notions of rationality, mostly minor variations of one another [140], the definition of intelligence is still debated and nebulous. There are several possibilities for such a definition. One approach is to define intelligence *to be* rationality, which is what is effectively done in [92] and [140]. From the perspective of artificial intelligence research, this is reasonable since rationality accurately describes the key characteristics of the agents that researchers want to build. At the other extreme, one could define intelligence to be what is measured by intelligence tests, as is done in [20]. This is actually quite useful, as it is generally agreed that everything measured by IQ tests is at least part of intelligence and so this approach provides a minimal definition of intelligence. So there are two possible approaches to defining intelligence: the narrow approach defines intelligence as essentially what is measured by IQ tests; the broad approach includes other cognitive abilities such as adapting to the environment, and showing wisdom and creativity [150]. Here, the narrow definition of intelligence is favoured. This has at least the advantage that intelligence can be measured, and leaves scope for separate investigation of the possible meanings of rationality, creativity, wisdom, and so on.

Furthermore, there is a psychological case for distinguishing rationality and intelligence. First, while the heuristics and biases research program showed that even highly intelligent people are prone to irrationality, *some* people are highly rational. Thus one might speculate from this that people with high intelligence (as measured by IQ tests) would be highly rational. Surprisingly, it turns out that rationality is only weakly correlated with intelligence [150]. There is now even a proposal for a rationality quotient test that evaluates those parts of rationality that are not covered by IQ tests [152]. Moreover, unlike intelligence, it turns out that rationality can be improved through training [68] and reflection [149].

In this book, rationality is taken to be the most central of the various cognitive abilities. Our artificial agents need to be rational because rationality is the best tool for solving problems and achieving appropriate goals. Think about some of the high points of the history of artificial intelligence: the chess-playing computer Deep Blue [77], the Jeopardy!-playing program Watson [49], the Go-playing program AlphaGo [145], and the poker-playing program Claudico [55]. These programs can beat virtually every human in the world. But they play very differently to humans. When Deep Blue beat Kasparov, it was widely noted how its extreme reliance on searching very large games trees was quite different to the way chess geniuses such as Kasparov played. But Deep Blue's behaviour was rational and effective, and even surprising on occasions in its choice of moves. AlphaGo showed such surprising behaviour as well, for example, with its famous move 37 in game 2 against Lee Sedol. In a subsequent version [146] of AlphaGo, the program called AlphaGo Zero learned to play Go to a superhuman level, easily beating AlphaGo itself, without any explicit human knowledge (although implicit human knowledge was used in the design of the program, for example). It learned solely by reinforcement learning, starting from random play without any supervision or use of human data. As noted in [146], "Humankind has accumulated Go knowledge from millions of games played over thousands of years, collectively distilled into patterns, proverbs and books. In the space of a few days, starting *tabula rasa*, AlphaGo Zero was able to rediscover much of this Go knowledge, as well as novel strategies that provide new insights into the oldest of games". For a popular account of these successes of AI and its history, see [173]. For a critical

account of the progress of AI, see [103]; in essence, other than for narrow, largely symbolic domains, existing AI systems lack genuine *understanding* of the world.

So artificial rational agents have already achieved world class standard in several narrow domains and this trend will undoubtedly continue. There is even progress on building *creative* artificial agents in the field of computational creativity [27, 104] which is a subfield of artificial intelligence. The topic of computers and creativity is studied in [16]. A more recent survey of this issue for artificial intelligence researchers is in [17]. Agents using artificial intelligence technology that produce music and art of sufficient standard to be played at concerts and exhibited have been produced. For example, [28] describes the underlying principles of computational creativity in music which have led to a large corpus of computer generated music that has been widely performed. Much of this music has been in the style of composers, such as Bach and Mozart, and has been so compelling that audiences have generally thought the computer generated music was actually music by the corresponding composer. In another creative domain, that of mathematics, programs are now widely used to find and verify proofs [165]. Also starting to appear are AI systems that provide strong guidance to the intuition of mathematicians and have led to the discovery of significant new theorems. See [35] and the references therein. One can reasonably anticipate that artificial mathematicians will be increasingly useful to human mathematicians as the century progesses, eventually to the point where artificial mathematicians are better at proofs than humans and also mathematically creative, so that human mathematicians will not be competitive without their artificial assistants. For a discussion of these issues by a Fields medallist, see [62, Section 2].

All this progress has been made with machines that have never shown the slightest indication of consciousness; indeed, it seems we are on a path to nonconscious, superintelligent machines, the implications of which are hard to glimpse at present. But note that all the progress so far has been in narrow domains, such as playing games, interpreting radiological scans, or composing music. While very impressive, these agents display nothing like the general-purpose intelligence that humans have. This is the most important and difficult, current issue in artificial intelligence research: how to build agents that can perform successfully in a wide variety of environments, as a high value of rationality, Υ, demands. Artificial general intelligence [164], the subfield of artificial intelligence research concerned with building general-purpose agents, addresses this issue. While agents that can perform well in a variety of environments are starting to make an appearance [110], artificial general intelligence has currently only barely started on what is likely to be decades of research to build such agents.

One can imagine that intelligence, creativity, and wisdom will *emerge* from rationality. Imagine the task of building an artificial mathematician using rationality as the basis for the agent architecture. A significant obstacle to achieving this is to define 'interestingness' of conjectures which is needed to define the performance measure of the agent. (Note that the methods of [35] are directly relevant to defining interestingness.) It will also be necessary to greatly improve heuristics for guiding proof search which motivates the actions the agent chooses. Indeed, mathematicians would argue that the hardest part of building such artificial mathematicians will be equipping them with methods for actually discovering beautiful and deep conjectures but, even with the conjectures, finding proofs of deep results is far beyond our current technology. Nevertheless, for the moment, concede the possibility of such an achievement. Then such an artificial mathematician would surely

be described as creative by humans, and intelligent as well, albeit in a narrow domain. Note, by the way, that, in contrast to rationality, no (narrow) definition of intelligence would be of any help in designing the architecture of the agent. More generally, it is the concept of rationality that drives the architecture of agents [140]. In other words, to rework the goal of DeepMind to 'solve intelligence', the problem is to *solve rationality*.

Artificial agents exploit the same rationality capabilities that humans have but eschew the irrational aspects of the human mind. Moreover, they leverage the huge computational and storage resources available to them in a way that humans simply cannot. This opens up a whole range of techniques for effective behaviour that are available to artifical agents but not to (unaugmented) humans. We are now reaching the point where, for many tasks, the methods of rational agents – the use of probability, logic, and decision theory, and the technologies, such as machine learning and vision processing, derived from these – are at least as effective as the methods available to the human mind. With the march of technology, it is clear that for an increasing range of tasks, humans are going to be outclassed and, eventually, by huge margins. For example, how long will it be before people regard as unreasonably dangerous being transported in a human-driven vehicle as compared with an autonomous vehicle?

The discussion above motivates the now mainstream idea that rationality rather than intelligence should be the foundational concept of the field of artificial intelligence. In fact, in retrospect, it would have been better if the field of artificial intelligence had instead been called *artificial rationality*, whose main objective is to build artificial rational agents.

The wider societal impacts of artificial rationality (AR) are now discussed. Every technology can be used for good or harm. AR is no exception to this; in fact, given the potential for this technology to end humankind, there is as much at stake for artificial rationality as there is for any other technology that we have developed, such as nuclear weapons and biotechnology. Even if one does not accept the possibility of superintelligent machines, there is still a lot at stake. AR is being used everyday, to assign credit ratings, to recommend purchases, to provide adaptive cruise control on motor vehicles, to provide automated video surveillance, and many other applications. Autonomous weapons systems have just started to be deployed. So, even in its currently limited applications, AR has the potential to do harm. It is clearly in our interests to make sure that AR is aligned with human values and is deployed only for the benefit of humanity. While there are dangers in pursuing the path we are on, from widespread unemployment [52], exploitation and inequality [29], to autonomous weapons [1], and to existential risk [9, 21, 58, 73], there are also potentially unlimited benefits and this is the outcome we must ensure [2, 139, 154, 162].

This issue of the potential for harm from AR technology has attracted considerable interest in the last decade and there are now a number of research institutes dedicated to what is commonly called 'robust and beneficial AI'. For example, the Future of Life Institute, founded in 2015 and based in Boston, funds research grants into robust and beneficial AI, has run an international conference Beneficial AI 2017, has published a set of principles (the Asilomar AI Principles) to guide AI research, and is concerned with other technology areas having similar potential dangers under the headings Biotech, Nuclear, and Climate. Institutes having similar goals include the Future of Humanity Institute based in Oxford, the Machine Intelligence Research Institute based in Berkeley, the Singularity University based in Moffett Field, California, and the Allen Institute for Artificial Intelligence based in Seattle.

Research into the risks of artificial intelligence is carried out in the important and currently highly active subfield of AI safety. For example, [5] and [139] provide discussions of research issues. An accessible account of the problem of control in AI can be found in [138]. The landscape of such AI research is presented in detail in [102], which categorizes the research topics as foundations, verification, validation, control, and security, to which one could add privacy, fairness, abuse, transparency, and policy from [5]. The issues of validation, verification, security and control, are now briefly described.

Validation is concerned with undesirable behaviours that can arise even if a system satisfies its specification. In other words, was the right system built? It can be very difficult to give a completely adequate specification of a complex system. Thus a system that is implemented correctly with respect to its specification may still exhibit surprising and undesirable behaviour. And there is no way to *prove* that the specification is completely adequate. The specification can, however, be used to prove properties of a correctly implemented system and these may be useful. Validation is hard even for conventional software; for an agent that is deployed in a complex, dynamic environment and is constantly modifying its behaviour through learning, it can become a wicked problem.

Verification is concerned with the correctness of the implementation with respect to the specification. In other words, was the system built right? Verification has been successfully used in a few cases in certain domains, but is hardly used at all in commercial software development. Generally, verifying AI software is much harder than this.

Security is concerned with attempts by external parties to modify the behaviour of an agent in undesirable ways. Cyber attacks are already a feature of daily life for everyone who accesses the Internet. Given that AI software may be responsible for ensuring human safety, for example, the software that controls an autonomous vehicle or the flight control software of an aircraft, the stakes here are already high and will increase as AI software becomes more autonomous.

Control is concerned with our ability to remain in control of AI as it becomes more capable. For example, even though an agent starts out with goals correctly aligned with human priorities, as it becomes more capable it may develop its own goals which are not so aligned. In the limit, this is the problem of a superintelligent agent that decides that humankind is a hindrance to its own goals and acts accordingly.

Epistemology [170] is the branch of philosophy concerned with the theory of knowledge. Here, the aspect of epistemology that is of most interest is that of empirical knowledge, the subject matter of empiricism [166]. In fact, the term empirical *belief* is preferred here since what is acquired from observations is belief that may be false in contrast with knowledge that has the connotation of being true. For an essay on beliefs from the perspective of an artificial intelligence researcher, see [125]. The ideas in the book could be useful to epistemologists in that it provides a precise definition of the concept of an empirical belief that has considerable generality and naturalness, and hence could be used to concretize epistemological theories. Furthermore, the approach of stochastic filtering, used here to acquire empirical beliefs, takes a particular philosophical position on belief acquisition that would be interesting to investigate. Finally, the highly expressive logic in which beliefs are expressed provides opportunities for investigations in formal epistemology.

A Gettier case is a scenario in which a justified, true belief is generally agreed by epistemologists not to be knowledge; see, for example, [80].

The term ‘schema’ comes from the psychologist Piaget [131] for whom a schema was

the basic building block of intelligent behavior – a way of organizing knowledge. More precisely, for Piaget a schema was "a cohesive, repeatable action sequence possessing component actions that are tightly interconnected and governed by a core meaning". For Piaget, schemas were procedural knowledge, whereas here they are declarative. Thus the schema concept introduced here is not a formalization of Piaget's concept, but does carry the same connotation of organized knowledge. The Haskell programming language is documented in [130].

Belief acquisition as envisaged in this book is a generalization of filtering that was originally invented in the field of signal processing [84] and later adopted in artificial intelligence, especially for robots [140, 156]. Linear dynamical systems and hidden Markov models are discussed, for example, in [14, 113, 140]. Particle filters are discussed in [30], [32], [113, 156], and, more briefly, in [14, 140].

Belief revision is discussed in [3, 69, 167]. Textbook treatments of machine learning can be found in [14, 113] and especially in the encyclopedic [115, 114]. The standard book on deep learning is [59]. For a discussion of the artificial intelligence tradition that "treats models of the world as primary, where learning is the process of model building", see [90].

The logic employed here has been greatly influenced by the higher-order logic in [6], but with the addition of modalities. Simple type theory was introduced in [26]. For an excellent discussion of the virtues of higher-order logic, [48] is highly recommended. Modal higher-order logic was also studied in [11, 50, 97, 116]. An extensive discussion of the reasoning capabilities of modal higher-order logic is presented in [101]. For discussion of modal beliefs for agents, see [47, 53, 105, 172], for example.

Chapter 2

State Distributions

This chapter is concerned with state distributions, the prototypical kind of empirical belief. First, action processes and observation processes are formally defined. Then the central concepts of an agent and an environment are introduced. Each of these are sequences of certain probability kernels that are regular conditional distributions, which provide correctness criteria for the definitions. This sets the stage for the definition of the concept of a state schema from which state distributions (in the form of probability measures or densities) are obtained by instantiating with the history. Acquisition of state distributions is by stochastic filtering, so the recurrence equations for filtering state distributions are presented. The proofs of these recurrence equations are based on transition and observation models being regular conditional distributions, which provides correctness criteria for these models. Later chapters will extend the results of this chapter from state distributions to more general empirical beliefs.

2.1 Action and Observation Processes

An agent operates in some environment, receiving observations from the environment, and performing actions on the environment in order to achieve some goal(s). The system containing the agent and the environment together can be modelled by a basic probability space $(\Omega, \mathfrak{S}, \mathsf{P})$. Intuitively, Ω is the set of all possible runs that the system may follow, where a run is the particular temporal sequence of events that occur. The probability measure P gives a distribution on all such runs. No further details about $(\Omega, \mathfrak{S}, \mathsf{P})$ are specified; instead, all attention is concentrated on several probability spaces that are induced by certain random variables.

To define these random variables, some preliminary concepts are needed.

Definition 2.1.1. An *action space* is a standard Borel space $(A, \mathcal{A})$. Each element of A is called an *action*.

Definition 2.1.2. An *observation space* is a standard Borel space $(O, \mathcal{O})$. Each element of O is called an *observation*.

It is not important at this stage to understand what a standard Borel space is, except that it is defined by a technical property satisfied by all the measure spaces one is ever

likely to meet in practice that ensures the existence of regular condition distributions, a concept that will be cucial for many of the key results in this book.

It will be convenient to introduce some notation.

Notation. Let $(A, \mathcal{A})$ be an action space and $(O, \mathcal{O})$ an observation space. For all $n \in \mathbb{N}_0$,

$$H_n \triangleq A \times O \times \cdots \times A \times O,$$

where there are n occurrences of A and n occurrences of O. H_0 is $\{()\}$. Each H_n becomes a measurable space with the usual product σ-algebra $\mathcal{H}_n \triangleq \mathcal{A} \otimes \mathcal{O} \otimes \cdots \otimes \mathcal{A} \otimes \mathcal{O}$, where there are n occurrences of $\mathcal{A}$ and n occurrences of $\mathcal{O}$.

Definition 2.1.3. Let $(A, \mathcal{A})$ be an action space and $(O, \mathcal{O})$ an observation space. *History space* is the set

$$H \triangleq \bigcup_{n \in \mathbb{N}_0} H_n.$$

An element of history space is called a *history*.

Each H_n is the set of histories of action-observation cycles up until the end of the nth cycle. Since $m \neq n$ implies $H_m \cap H_n = \emptyset$, it follows that $H = \coprod_{n \in \mathbb{N}_0} H_n$. H can be made into a measurable space in the obvious way: define $\mathcal{H} \triangleq \bigoplus_{n \in \mathbb{N}_0} \mathcal{H}_n$. Then $(H, \mathcal{H})$ is a measurable space.

Notation. $H_\infty \triangleq A \times O \times A \times O \times \cdots$.

Definition 2.1.4. Let $(A, \mathcal{A})$ be an action space and $(O, \mathcal{O})$ an observation space. Then *interaction space* is the measurable space $(H_\infty, \mathcal{H}_\infty)$, where $\mathcal{H}_\infty$ is the usual product σ-algebra on H_∞.

Each element of H_∞ is called an *interaction sequence*.

Note that interaction space, being a countable product of standard Borel spaces, is a standard Borel space.

Two central objects of study are action processes and observation processes.

Definition 2.1.5. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(A, \mathcal{A})$ an action space. An *action process* (*based on* A) is a stochastic process

$$\mathbf{a} : \Omega \to A^{\mathbb{N}}.$$

Definition 2.1.6. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(O, \mathcal{O})$ an observation space. An *observation process* (*based on* O) is a stochastic process

$$\mathbf{o} : \Omega \to O^{\mathbb{N}}.$$

Notation. For all $n \in \mathbb{N}$, $\mathbf{a}_n : \Omega \to A$ is defined by $\mathbf{a}_n(\omega) = \mathbf{a}(\omega)(n)$, for all $\omega \in \Omega$.

For all $n \in \mathbb{N}$, $\mathbf{o}_n : \Omega \to O$ is defined by $\mathbf{o}_n(\omega) = \mathbf{o}(\omega)(n)$, for all $\omega \in \Omega$.

Each $\mathbf{a}_n$ and $\mathbf{o}_n$ is clearly a random variable.

Action and observation processes can be combined into interaction processes.

Definition 2.1.7. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $\mathbf{a}$ an action process based on A, and $\mathbf{o}$ an observation process based on O. Then the *interaction process*

$$\boldsymbol{\iota} : \Omega \to H_\infty$$

(*for* $\mathbf{a}$ *and* $\mathbf{o}$) is defined by $\boldsymbol{\iota} = (\mathbf{a}_1, \mathbf{o}_1, \mathbf{a}_2, \mathbf{o}_2, \ldots)$.

Of course, it is just as easy to start with interaction processes, and derive action and observation processes.

Random variables of the form $(\mathbf{a}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{o}_n) : \Omega \to H_n$ appear often below. Thus, for notational convenience, a shorthand is introduced.

Notation. Let $(A, \mathcal{A})$ be an action space, $(O, \mathcal{O})$ an observation space, $\mathbf{a}$ an action process based on A, and $\mathbf{o}$ an observation process based on O. For all $n \in \mathbb{N}_0$,

$$\mathbf{h}_n \triangleq (\mathbf{a}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{o}_n).$$

Note that, for all $\omega \in \Omega$, $\mathbf{h}_n(\omega) \in H_n$ is the history up to the end of n action-observation cycles. In particular, $\mathbf{h}_0 : \Omega \to H_0$ satisfies $\mathbf{h}_0(\omega) = ()$, for all $\omega \in \Omega$. Also $(\mathbf{h}_{n-1}, \mathbf{a}_n) : \Omega \to H_{n-1} \times A$, where $(\mathbf{h}_{n-1}, \mathbf{a}_n) = (\mathbf{a}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n)$, is a random variable.

2.2 Agents and Environments

Next the fundamental concepts of an agent and an environment are introduced.

Definition 2.2.1. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, and $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process. An *agent* (*for* $\mathbf{a}$ *and* $\mathbf{o}$) is a sequence $\Lambda \triangleq (\Lambda_n)_{n \in \mathbb{N}}$, where

$$\Lambda_n : H_{n-1} \to \mathcal{P}(A)$$

is a regular conditional distribution of $\mathbf{a}_n$ given $\mathbf{h}_{n-1}$, for all $n \in \mathbb{N}$.

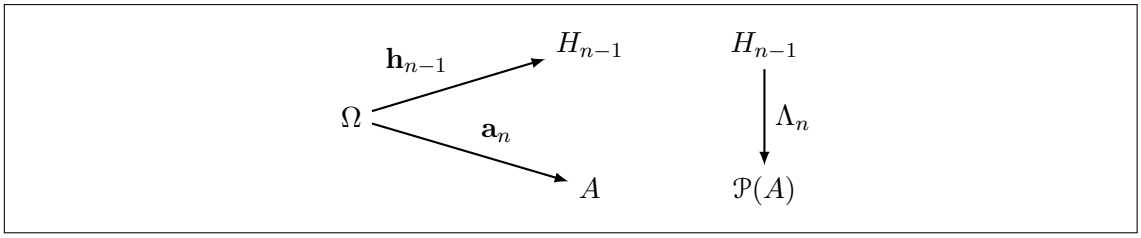

Figure 2.1: A component Λ_n of an agent for an action process and an observation process

In other words, for all $n \in \mathbb{N}$, Λ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{a}_n^{-1}(C) \mid \mathbf{h}_{n-1}) = \lambda\omega.\Lambda_n(\mathbf{h}_{n-1}(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{A}$. Thus Λ_n is not just any probability kernel; it must also be an appropriate regular conditional distribution that is 'consistent' with $\mathbf{a}$ and $\mathbf{o}$. According to Proposition 2.2.1 below, each Λ_n exists and is unique $\mathcal{L}(\mathbf{h}_{n-1})$-a.e., so that it is possible to refer to *the* agent (for $\mathbf{a}$ and $\mathbf{o}$). An agent takes as input the history so far up to the last observation generated by the environment and returns a distribution on the actions that could be performed by the agent.

The expression $\mathsf{P}(\mathbf{a}_n^{-1}(C) \mid \mathbf{h}_{n-1})$ is a conditional probability that has a distinctly subtle definition due to Kolmogorov [89] and has become a fundamental concept of modern probability theory. For some purposes it is sufficient to rely on its intuitive meaning: 'the conditional probability of the event $\mathbf{a}_n^{-1}(C)$ given that the value of $\mathbf{h}_{n-1}$ is known'. The probability kernel Λ_n provides a kind of representation of the conditional probability that is much more convenient to use in practice than the conditional probability itself.

Definition 2.2.2. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, and $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process. An *environment* (*for* $\mathbf{a}$ *and* $\mathbf{o}$) is a sequence $\Xi \triangleq (\Xi_n)_{n\in\mathbb{N}}$, where

$$\Xi_n : H_{n-1} \times A \to \mathcal{P}(O)$$

is a regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n)$, for all $n \in \mathbb{N}$.

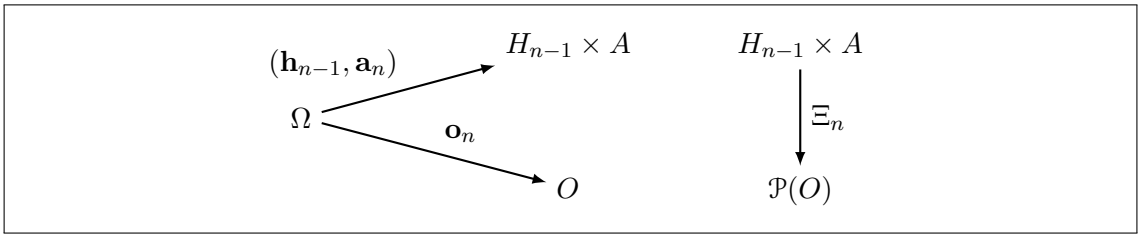

Figure 2.2: A component Ξ_n of an environment for an action process and an observation process

In other words, for all $n \in \mathbb{N}$, Ξ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{o}_n^{-1}(D) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n)) = \lambda\omega.\Xi_n((\mathbf{h}_{n-1}, \mathbf{a}_n)(\omega))(D) \text{ a.s.},$$

for all $D \in \mathcal{O}$. According to Proposition 2.2.1, each Ξ_n exists and is unique $\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n))$-a.e., so that it is possible to refer to *the* environment (for $\mathbf{a}$ and $\mathbf{o}$). Thus Ξ_n is not just any probability kernel; it must also be an appropriate regular conditional distribution that is 'consistent' with $\mathbf{a}$ and $\mathbf{o}$. An environment takes as input the history so far up to the last action performed by the agent and returns a distribution on the observations that could be generated by the environment.

Example 2.2.1. Note that

$$\Lambda_1 : \mathcal{P}(A)$$
$$\Lambda_1 \otimes \Xi_1 : \mathcal{P}(H_1)$$
$$\Lambda_1 \otimes \Xi_1 \otimes \Lambda_2 : \mathcal{P}(H_1 \times A)$$
$$\Lambda_1 \otimes \Xi_1 \otimes \Lambda_2 \otimes \Xi_2 : \mathcal{P}(H_2)$$

and so on.

More generally, by Proposition A.7.5,

$$\Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_{n-1} \otimes \Xi_{n-1} \otimes \Lambda_n : \mathcal{P}(H_{n-1} \times A)$$

and

$$\Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_n \otimes \Xi_n : \mathcal{P}(H_n).$$

Figure 2.3 illustrates the dependency graph for (the first five time steps of) an interaction process. It is convenient to denote each vertex by the random variable that generates the σ-algebra that labels the vertex. Let the topological order of the vertices in the dependency graph be $\mathbf{a}_1$, $\mathbf{o}_1$, $\mathbf{a}_2$, $\mathbf{o}_2$, and so on. The edges indicate the direct dependencies. In the graph, every predecessor of a vertex is assumed to be a parent of the vertex. Thus the Markov property of Definition A.6.5 is satisfied, no matter what the σ-algebras $\mathcal{F}_i$ are. In general, fewer edges may be needed; in any case, for the dependency graph to correctly model the application at hand, it must be Markov.

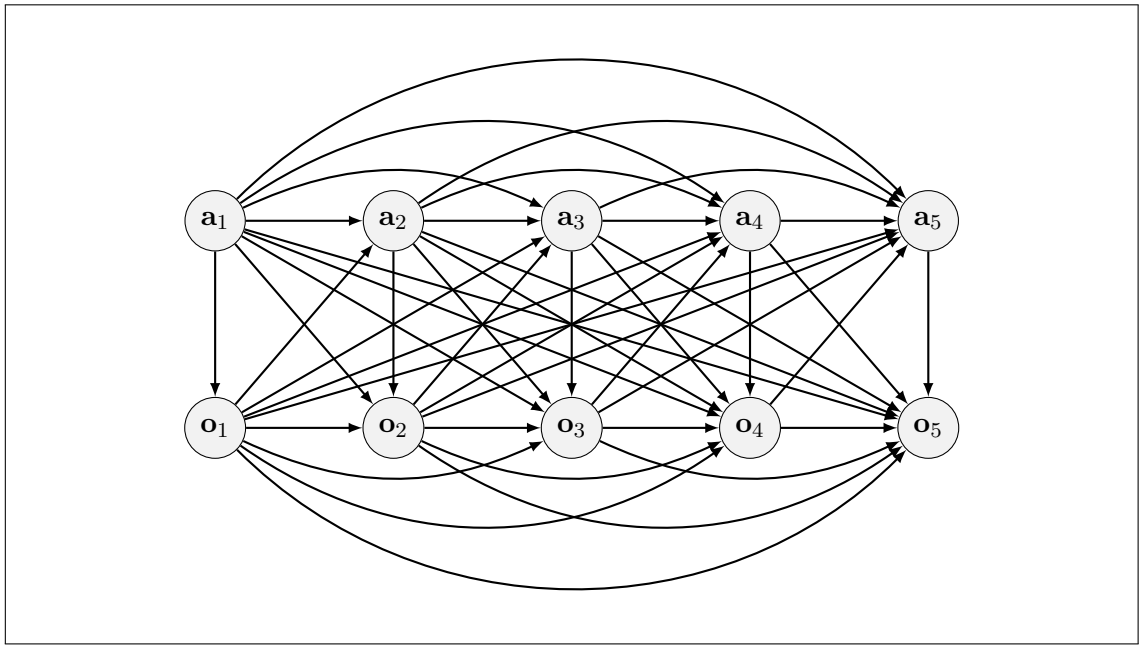

Figure 2.3: Dependency graph for first 5 time steps of an interaction process

Now it is proved that each action and observation process together determine an agent and an environment that are essentially unique.

Proposition 2.2.1. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, and $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process. Then there exists an agent Λ and an environment Ξ for $\mathbf{a}$ and $\mathbf{o}$, where Λ_n is unique $\mathcal{L}(\mathbf{h}_{n-1})$-a.e. and Ξ_n is unique $\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n))$-a.e., for all $n \in \mathbb{N}$. Furthermore, for all $n \in \mathbb{N}$,*

$$\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n)) = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_{n-1} \otimes \Xi_{n-1} \otimes \Lambda_n$$

and

$$\mathcal{L}(\mathbf{h}_n) = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_n \otimes \Xi_n.$$

Proof. According to Proposition A.5.16, since A is a standard Borel space, each Λ_n exists and is unique $\mathcal{L}(\mathbf{h}_{n-1})$-a.e. Similarly, each Ξ_n exists and is unique $\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n))$-a.e.

By Proposition A.7.13, for all $E \in \mathcal{A} \otimes \mathcal{O} \otimes \cdots \otimes \mathcal{A} \otimes \mathcal{O} \otimes \mathcal{A}$,

$$\mathsf{P}((\mathbf{h}_{n-1}, \mathbf{a}_n)^{-1}(E)) = (\Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_{n-1} \otimes \Xi_{n-1} \otimes \Lambda_n)(E).$$

That is,

$$\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n)) = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_{n-1} \otimes \Xi_{n-1} \otimes \Lambda_n.$$

The other part is similar. □

Since each Λ_n is unique $\mathcal{L}(\mathbf{h}_{n-1})$-a.e., it is possible to refer to *the* agent for an action process and an observation process. Similarly, it is possible to refer to *the* environment for an action process and an observation process.

The preceding definitions of an agent and an environment provide an external view of the concepts of an agent and an environment. All that is apparent to an external observer about an agent is the sequence of actions computed by the agent given the history so far (and analogously for the environment). This view motivates the definition of an agent given above and is what was called the abstract view of an agent in Section 1.3. However, much of the book is concerned with the internal view of an agent for which the details of its architecture are important; this is called the architectural view of an agent. The abstract agent can be regarded as the policy of the architectural agent, the policy being the function that is used to choose actions.

Agents as defined here are stochastic in the sense that the actual action selected by the agent needs to be sampled from the distribution on A given by the agent policy. Stochastic agents can be useful for some applications even when a deterministic agent would be possible, since they avoid being predictable. In addition, a stochastic agent can be used to balance exploration versus exploitation [140, Section 21.3]. For exploration, the agent uses a distribution on actions that has a bigger variance; sampling from such 'wider' distributions produces a bigger range of actions that enable exploration. For exploitation, the agent uses a distribution that has a smaller variance.

For each $n \in \mathbb{N}$, Λ_n is a distinct function because its domain H_{n-1} varies with n. From a practical point of view this is not convenient. However, it is simple to reduce the sequence of action functions to a single function: define $\overline{\Lambda} : \bigcup_{n \in \mathbb{N}_0} H_n \to \mathcal{P}(A)$ by $\overline{\Lambda}(h) = \Lambda_n(h)$, whenever $h \in H_n$. Note that, for all $n \in \mathbb{N}$, $\Lambda_n = \overline{\Lambda} \circ inj_n$, where $inj_n : H_{n-1} \to \bigcup_{m \in \mathbb{N}_0} H_m$ is the canonical injection. An agent system would store (and maintain and learn) the function $\overline{\Lambda}$. Analogous remarks apply to $(\Xi_n)_{n \in \mathbb{N}}$.

The next result provides the basis for simulating any choice of agent and environment.

Proposition 2.2.2. *(Simulation) Let $(A, \mathcal{A})$ be an action space, $(O, \mathcal{O})$ an observation space, and $(\Lambda_n : H_{n-1} \to \mathcal{P}(A))_{n \in \mathbb{N}}$ and $(\Xi_n : H_{n-1} \times A \to \mathcal{P}(O))_{n \in \mathbb{N}}$ sequences of probability kernels. Then there exists a probability space $(\Omega, \mathfrak{S}, \mathsf{P})$, an action process $\mathbf{a} : \Omega \to A^{\mathbb{N}}$, and an observation process $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ such that*

$$\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n)) = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_{n-1} \otimes \Xi_{n-1} \otimes \Lambda_n$$

and

$$\mathcal{L}(\mathbf{h}_n) = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_n \otimes \Xi_n,$$

for all $n \in \mathbb{N}$. *Furthermore,* $(\Lambda_n)_{n\in\mathbb{N}}$ *is an agent for* $\mathbf{a}$ *and* $\mathbf{o}$ *and* $(\Xi_n)_{n\in\mathbb{N}}$ *is an environment for* $\mathbf{a}$ *and* $\mathbf{o}$.

Proof. By Proposition A.8.1, there exists a unique probability measure κ on $(H_\infty, \mathcal{H}_\infty)$ such that

$$\kappa \circ \pi^{-1}_{1,\ldots,2n-1} = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_{n-1} \otimes \Xi_{n-1} \otimes \Lambda_n$$

and

$$\kappa \circ \pi^{-1}_{1,\ldots,2n} = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_n \otimes \Xi_n,$$

where $\pi_{1,\ldots,2n-1} : H_\infty \to H_{n-1} \times A$ and $\pi_{1,\ldots,2n} : H_\infty \to H_n$ are the canonical projections. Let $\Omega \triangleq H_\infty$, $\mathfrak{S} \triangleq \mathcal{H}_\infty$, and $\mathsf{P} \triangleq \kappa$. Thus $(\Omega, \mathfrak{S}, \mathsf{P})$ is a probability space. Define $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ to be the canonical projection. Hence $\mathbf{a}$ is an action process based on A. Similarly, define $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ to be the canonical projection. Hence $\mathbf{o}$ is an observation process based on O. Now, since $(\mathbf{h}_{n-1}, \mathbf{a}_n) = \pi_{1,\ldots,2n-1}$,

$$\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n)) = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_{n-1} \otimes \Xi_{n-1} \otimes \Lambda_n,$$

for all $n \in \mathbb{N}$. Also, since $\mathbf{h}_n = \pi_{1,\ldots,2n}$,

$$\mathcal{L}(\mathbf{h}_n) = \Lambda_1 \otimes \Xi_1 \otimes \cdots \otimes \Lambda_n \otimes \Xi_n,$$

for all $n \in \mathbb{N}$.

Finally, it needs to be shown that Λ is an agent for $\mathbf{a}$ and $\mathbf{o}$ and Ξ is an environment for $\mathbf{a}$ and $\mathbf{o}$. By Proposition A.8.2, for all $n \in \mathbb{N}$ and $C \in \mathcal{A}$, P-almost surely, $\mathsf{P}(\mathbf{a}_n^{-1}(C) \mid \mathbf{h}_{n-1}) = \lambda\omega.\Lambda_n(\mathbf{h}_{n-1}(\omega))(C)$. Hence Λ is an agent for $\mathbf{a}$ and $\mathbf{o}$. Similarly, by Proposition A.8.2, for all $n \in \mathbb{N}$ and $D \in \mathcal{O}$, P-almost surely, $\mathsf{P}(\mathbf{o}_n^{-1}(D) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n)) = \lambda\omega.\Xi_n((\mathbf{h}_{n-1}, \mathbf{a}_n)(\omega))(D)$. Hence Ξ is an environment for $\mathbf{o}$ for $\mathbf{a}$ and $\mathbf{o}$. □

The definitions of $(\Lambda_n)_{n\in\mathbb{N}}$ and $(\Xi_n)_{n\in\mathbb{N}}$ in Proposition 2.2.2 can be completely arbitrary. The result then shows that there exist a probability space, an action process $\mathbf{a}$, and an observation process $\mathbf{o}$ such that $(\Lambda_n)_{n\in\mathbb{N}}$ is an agent and $(\Xi_n)_{n\in\mathbb{N}}$ is an environment for $\mathbf{a}$ and $\mathbf{o}$. Proposition 2.2.2 provides the theoretical foundation for the simulation of an agent-environment system for any desired agent and environment.

2.3 State Schemas

In this section, state distributions, the prototypical kind of empirical belief, are studied. Intuitively, a state models information about the underlying state of the environment in which the agent is deployed that can be used by the agent to select actions. However, the state is usually not observable by the agent; instead the agent receives observations that are probabilistically dependent on the state. In the literature, this setting for agents and environments is called a partially observable Markov decision process. Also, since states

can have arbitrary structure, the setting here also includes that of dynamic Bayesian networks, which can be thought of as the case where the state is a product space and there is just a single action that can be thought of as an exogenous transition whose definition takes advantage of the sparseness of the underlying Bayesian network, as does the observation model. The central concept is that of a state schema from which state distributions can be obtained by instantiating with the history.

Here are the ingredients additional to action and observation processes needed for this setting. These additional ingredients are states, state processes, state schemas, observation models, and transition models.

Definition 2.3.1. A *state space* is a standard Borel space $(S, \mathcal{S})$. Each element of S is called an *state*.

Definition 2.3.2. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(S, \mathcal{S})$ a state space. A *state process* is a stochastic process

$$\mathbf{s} : \Omega \to S^{\mathbb{N}_0}.$$

Notation. For all $n \in \mathbb{N}_0$, $\mathbf{s}_n : \Omega \to S$ is defined by $\mathbf{s}_n(\omega) = \mathbf{s}(\omega)(n)$, for all $\omega \in \Omega$.

Each $\mathbf{s}_n$ is clearly a random variable. Note that the indexing for a state process starts from 0.

The introduction of state allows for some strong but realistic conditional independence assumptions to be made.

Definition 2.3.3. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(S, \mathcal{S})$ a state space, $(O, \mathcal{O})$ an observation space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ a state process, and $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process. For the dependency graph illustrated in Figure 2.4, consider the topological order of vertices $\mathbf{s}_0$, $\mathbf{a}_1$, $\mathbf{s}_1$, $\mathbf{o}_1$, $\mathbf{a}_2$, $\mathbf{s}_2$, $\mathbf{o}_2$, $\ldots$. The stochastic process

$$(\mathbf{a}, \mathbf{s}, \mathbf{o}) : \Omega \to A^{\mathbb{N}} \times S^{\mathbb{N}_0} \times O^{\mathbb{N}}$$

is *Markov* if it satisfies the Markov property with respect to this dependency graph and topological order.

More explicitly, the process $(\mathbf{a}, \mathbf{s}, \mathbf{o})$ is Markov if

$$\sigma(\mathbf{a}_n) \underset{\sigma(\mathbf{a}_1,\ldots,\mathbf{a}_{n-1},\mathbf{o}_1,\ldots,\mathbf{o}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{a}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{s}_0, \ldots, \mathbf{s}_{n-1}, \mathbf{o}_1, \ldots, \mathbf{o}_{n-1}),$$
$$\sigma(\mathbf{s}_n) \underset{\sigma(\mathbf{a}_n,\mathbf{s}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{s}_0, \ldots, \mathbf{s}_{n-1}, \mathbf{o}_1, \ldots, \mathbf{o}_{n-1}),$$

and

$$\sigma(\mathbf{o}_n) \underset{\sigma(\mathbf{s}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{s}_0, \ldots, \mathbf{s}_n, \mathbf{o}_1, \ldots, \mathbf{o}_{n-1}),$$

for all $n \in \mathbb{N}$. A process $(\mathbf{a}, \mathbf{s}, \mathbf{o})$ that is Markov in this sense is a partially observable Markov decision process, in the usual terminology.

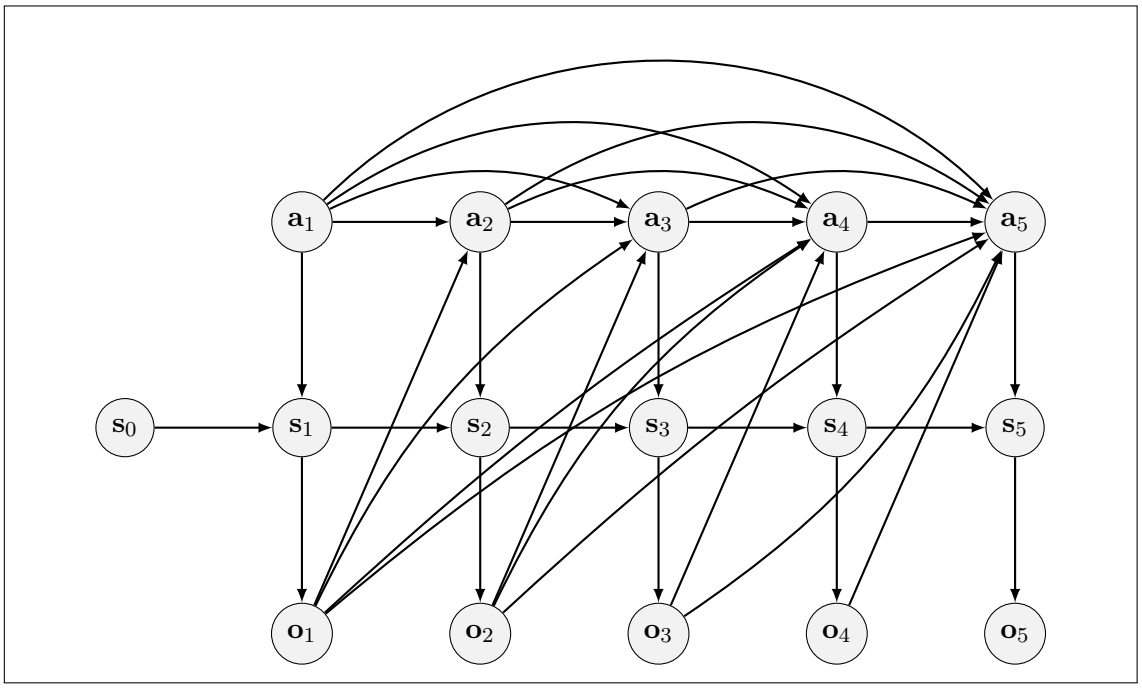

Figure 2.4: Dependency graph for first 5 time steps of a stochastic process $(\mathbf{a}, \mathbf{s}, \mathbf{o})$

Note. The conditional independence property for $\mathbf{s}_0$ is

$$\sigma(\mathbf{s}_0) \underset{\{\emptyset,\Omega\}}{\perp\!\!\!\perp} \{\emptyset, \Omega\},$$

which is trivially true. Also the conditional independence property for $\mathbf{a}_1$,

$$\sigma(\mathbf{a}_1) \underset{\{\emptyset,\Omega\}}{\perp\!\!\!\perp} \sigma(\mathbf{s}_0),$$

is equivalent to the requirement that $\sigma(\mathbf{a}_1)$ and $\sigma(\mathbf{s}_0)$ are independent.

An important task at each time step for an agent is to determine the distribution on the current state given the history up to the current time, since the distribution on the current state is needed to choose the next action. (See the remark on this below.) The task of determining the distribution on the current state is known as *filtering*. To introduce filtering, the concept of a state schema is needed.

Definition 2.3.4. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(S, \mathcal{S})$ a state space, $(O, \mathcal{O})$ an observation space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ a state process, and $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process. A *state schema* is a sequence $\mu \triangleq (\mu_n)_{n\in\mathbb{N}_0}$, where

$$\mu_n : H_n \to \mathcal{P}(S)$$

is a regular conditional distribution of $\mathbf{s}_n$ given $\mathbf{h}_n$, for all $n \in \mathbb{N}_0$.

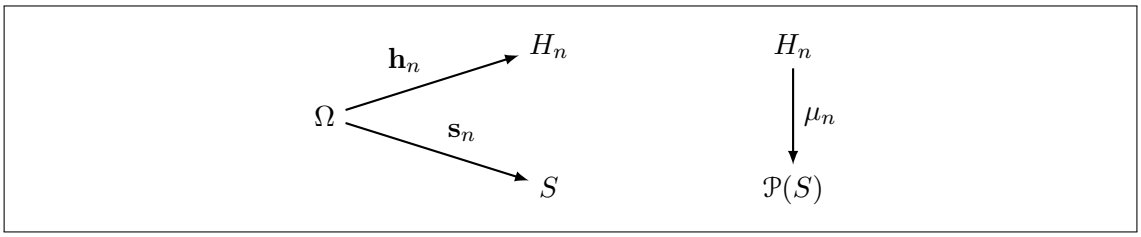

Figure 2.5: A component μ_n of a state schema

In other words, for all $n \in \mathbb{N}_0$, μ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{S}$. According to Proposition A.5.16, each μ_n exists and is unique $\mathcal{L}(\mathbf{h}_n)$-a.e., so that it is possible to refer to *the* state schema. A state schema takes as input the history so far and returns a distribution on the possible states. The concept of a state schema is a special case of the key concept of a schema that will be introduced in Chapter 3.

It is assumed that the initial schema $\mu_0 : H_0 \to \mathcal{P}(S)$ is known to the agent. Since H_0 is the singleton set $\{()\}$, μ_0 can be identified with a distribution on S, that is, $\mu_0 : \mathcal{P}(S)$.

Formally, the next action is chosen with the agent function $\Lambda_n : H_{n-1} \to \mathcal{P}(A)$. However, it is commonly assumed that this function 'factors' through $\mathcal{P}(S)$, so that, if the distribution on the current state is known, then the next action can be chosen using a function having signature $S \to \mathcal{P}(A)$. Towards this, suppose that $\alpha : S \to \mathcal{P}(A)$ is a suitable function for selecting actions depending on the current state. Then, for $h_{n-1} \in H_{n-1}$, one has $\mu_{n-1}(h_{n-1}) : \mathcal{P}(S)$ and so the fusion

$$\mu_{n-1}(h_{n-1}) \odot \alpha : \mathcal{P}(A)$$

can be sampled to select an action given the current history h_{n-1}.

Some motivation for the requirement that state schema components be regular conditional distributions may be helpful. There are many probability kernels having signature $H_n \to \mathcal{P}(S)$, but (essentially) only one is of interest here and that is the one that captures the probabilistic relationship between the random variables $\mathbf{h}_n$ and $\mathbf{s}_n$ given by the conditional probability $\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n)$, for $C \in \mathcal{S}$. This is the regular conditional distribution μ_n that represents $\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n)$ via the condition $\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C)$ a.s., for all $C \in \mathcal{S}$, and thus provides a correctness criterion for (components of) state schemas. Identifying exactly which probability kernels having signature $H_n \to \mathcal{P}(S)$ should be (components of) state schemas turns out to be important; for example, see the comments about the proof of Proposition 2.3.2 following Figure2.9 below. Analogous remarks apply to the definitions of transition models and observation models below.

Now the transition model and observation model, which are needed for filtering, are introduced. (More general kinds of transition and observation models will be defined in Chapter 4.)

Definition 2.3.5. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(S, \mathcal{S})$ a state space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, and $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ a state process. A *transition model* is a sequence $\tau \triangleq (\tau_n)_{n\in\mathbb{N}}$, where

$$\tau_n : A \times S \to \mathcal{P}(S)$$

is a regular conditional distribution of $\mathbf{s}_n$ given $(\mathbf{a}_n, \mathbf{s}_{n-1})$, for all $n \in \mathbb{N}$.

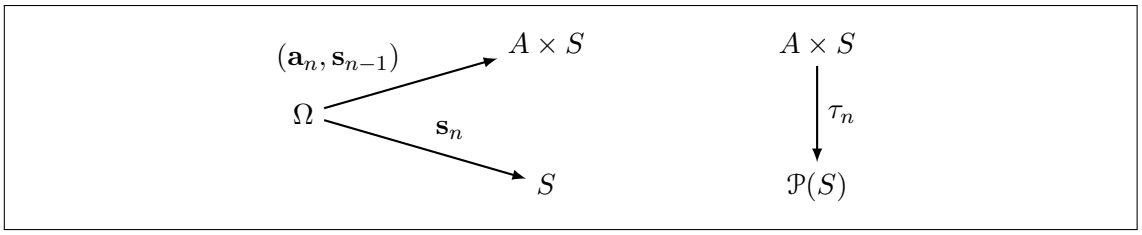

Figure 2.6: A component τ_n of a transition model

In other words, for all $n \in \mathbb{N}$, τ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{a}_n, \mathbf{s}_{n-1})) = \lambda\omega.\tau_n((\mathbf{a}_n, \mathbf{s}_{n-1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{S}$. According to Proposition A.5.16, for each $n \in \mathbb{N}$, such a τ_n exists and is unique $\mathcal{L}((\mathbf{a}_n, \mathbf{s}_{n-1}))$-a.e. A transition model takes as input the current state and action and returns a distribution on the states that could result from the transition.

Although the concept of a transition model has been defined for arbitary processes, the concept is mostly useful for processes that are Markov since they satisfy the second part

$$\sigma(\mathbf{s}_n) \underset{\sigma(\mathbf{a}_n, \mathbf{s}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{s}_1, \ldots, \mathbf{s}_{n-1}, \mathbf{o}_1, \ldots, \mathbf{o}_{n-1})$$

of the Markov property. This states intuitively that the next state depends only on the current state and the next action, but not on any action, observation, or state before the current state. If the Markov property is satisfied, then, for all $n \in \mathbb{N}$,

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{s}_1, \ldots, \mathbf{s}_{n-1}, \mathbf{o}_1, \ldots, \mathbf{o}_{n-1})) = \lambda\omega.\tau_n((\mathbf{a}_n, \mathbf{s}_{n-1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{S}$.

Next comes the concept of an observation model.

Definition 2.3.6. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(S, \mathcal{S})$ a state space, $(O, \mathcal{O})$ an observation space, $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ a state process, and $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process. An *observation model* is a sequence $\xi \triangleq (\xi_n)_{n \in \mathbb{N}}$, where

$$\xi_n : S \to \mathcal{P}(O)$$

is a regular conditional distribution of $\mathbf{o}_n$ given $\mathbf{s}_n$, for all $n \in \mathbb{N}$.

In other words, for all $n \in \mathbb{N}$, ξ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid \mathbf{s}_n) = \lambda\omega.\xi_n(\mathbf{s}_n(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{O}$. According to Proposition A.5.16, for each $n \in \mathbb{N}$, such a ξ_n exists and is unique $\mathcal{L}(\mathbf{s}_n)$-a.e. An observation model takes as input the current state and returns a distribution on the observations that could be generated by the environment.

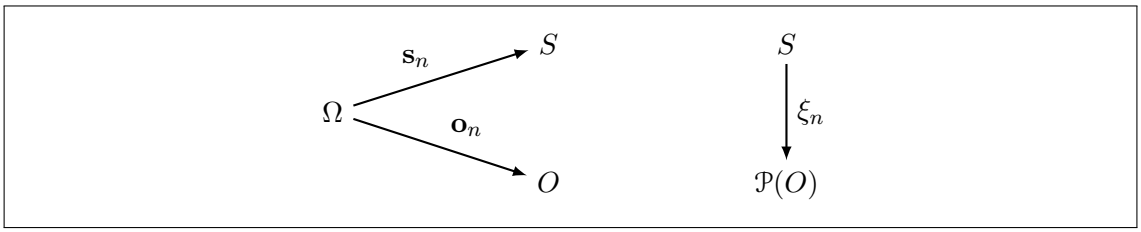

Figure 2.7: A component ξ_n of an observation model

Although the concept of an observation model has been defined for arbitary processes, the concept is mostly useful for processes that are Markov since they satisfy the first part

$$\sigma(\mathbf{o}_n) \underset{\sigma(\mathbf{s}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{s}_1, \ldots, \mathbf{s}_n, \mathbf{o}_1, \ldots, \mathbf{o}_{n-1})$$

of the Markov property. This states intuitively that the next observation depends only on the current state, but not on any action, observation, or state before the current state. If the Markov property is satisfied, then, for all $n \in \mathbb{N}$,

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{s}_1, \ldots, \mathbf{s}_n, \mathbf{o}_1, \ldots, \mathbf{o}_{n-1})) = \lambda\omega.\xi_n(\mathbf{s}_n(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{O}$.

If $(\mu_n : H_n \to \mathcal{P}(S))_{n\in\mathbb{N}_0}$ is a state schema, then $\mu_n(h_n) : \mathcal{P}(S)$, for some $h_n \in H_n$, is called a *state distribution* (*in probability measure form*). Similarly, let υ_S be a σ-finite measure on $\mathcal{S}$ and $\breve{\mu}_n : H_n \to \mathcal{D}(S)$ a conditional density such that $\mu_n = \breve{\mu}_n \cdot \upsilon_S$. Then then $\breve{\mu}_n(h_n) : \mathcal{D}(S)$, for some $h_n \in H_n$, is called a *state distribution* (*in density form*).

From a theoretical point of view, the primary objects of interest in this chapter are the conditional probabilities $\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n)$, $\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{a}_n, \mathbf{s}_{n-1}))$, and $\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid \mathbf{s}_n)$. However, conditional probabilities are not convenient to deal with practically; instead, one passes to the corresponding regular conditional distributions, given by Proposition A.5.16, that are more convenient being probability kernels. In the subsequent theoretical development, it will be the case that primarily it is the properties of conditional probabilities that are used to prove many results and that one converts between regular conditional distributions and conditional probabilities as necessary to exploit these properties. An example of this methodology appears in Proposition 2.3.2 below.

To determine the next state distribution, the current state distribution and a recurrence equation that determines the next distribution from the current one are used. The next main result (Proposition 2.3.2) establishes this recurrence equation in an explicit fashion. (More general results for arbitrary schemas and empirical beliefs will be proved in Chapter 4.) To prove the result about the recurrence equation, some consequences of the Markov assumption need to be established first.

Proposition 2.3.1. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(S, \mathcal{S})$ a state space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ a state process, $(\mu_n : H_n \to \mathcal{P}(S))_{n\in\mathbb{N}_0}$ the state schema, $(\tau_n : A \times S \to \mathcal{P}(S))_{n\in\mathbb{N}}$ the transition model, and $(\xi_n : S \to \mathcal{P}(O))_{n\in\mathbb{N}}$ the observation model. Suppose that the stochastic process $(\mathbf{a}, \mathbf{s}, \mathbf{o}) : \Omega \to A^{\mathbb{N}} \times S^{\mathbb{N}_0} \times O^{\mathbb{N}}$ is Markov. Then the following hold.*

1. $\lambda(h,a).\mu_n(h) : H_n \times A \to \mathcal{P}(S)$ *is a regular conditional distribution of* $\mathbf{s}_n$ *given* $(\mathbf{h}_n, \mathbf{a}_{n+1})$*, for all* $n \in \mathbb{N}_0$.
2. $\lambda(h,a,s).\tau_n(a,s) : H_{n-1} \times A \times S \to \mathcal{P}(S)$ *is a regular conditional distribution of* $\mathbf{s}_n$ *given* $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_{n-1})$*, for all* $n \in \mathbb{N}$.
3. $\lambda(h,a,s).\xi_n(s) : H_{n-1} \times A \times S \to \mathcal{P}(O)$ *is a regular conditional distribution of* $\mathbf{o}_n$ *given* $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_n)$*, for all* $n \in \mathbb{N}$.

Proof. 1. For all $n \in \mathbb{N}_0$, since μ_n is a regular conditional distribution of $\mathbf{s}_n$ given $\mathbf{h}_n$,

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{S}$. By the Markov property,

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{s}_n),$$

and so, by Proposition A.6.1,

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n) = \mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) \text{ a.s.},$$

for all $C \in \mathcal{S}$. Hence

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C) \text{ a.s.},$$

that is,

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.(\lambda(h,a).\mu_n(h))((\mathbf{h}_n, \mathbf{a}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{S}$. In other words, $\lambda(h,a).\mu_n(h) : H_n \times A \to \mathcal{P}(S)$ is a regular conditional distribution of $\mathbf{s}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$.

2. For all $n \in \mathbb{N}$, since τ_n is a regular conditional distribution of $\mathbf{s}_n$ given $(\mathbf{a}_n, \mathbf{s}_{n-1})$,

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{a}_n, \mathbf{s}_{n-1})) = \lambda\omega.\tau_n((\mathbf{a}_n, \mathbf{s}_{n-1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{S}$. By the Markov property,

$$\sigma(\mathbf{h}_{n-1}) \underset{\sigma(\mathbf{a}_n, \mathbf{s}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{s}_n)$$

and so, by Proposition A.6.1,

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{a}_n, \mathbf{s}_{n-1})) = \mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_{n-1})) \text{ a.s.},$$

for all $C \in \mathcal{S}$. Hence

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_{n-1})) = \lambda\omega.\tau_n((\mathbf{a}_n, \mathbf{s}_{n-1})(\omega))(C) \text{ a.s.},$$

that is,

$$\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_{n-1})) = \lambda\omega.(\lambda(h,a,s).\tau_n(a,s))((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_{n-1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{S}$. In other words, $\lambda(h,a,s).\tau_n(a,s)$ is a regular conditional distribution of $\mathbf{s}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_{n-1})$, for all $n \in \mathbb{N}$.

3. For all $n \in \mathbb{N}$, since ξ_n is a regular conditional distribution of $\mathbf{o}_n$ given $\mathbf{s}_n$,

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid \mathbf{s}_n) = \lambda\omega.\xi_n(\mathbf{s}_n(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{O}$. By the Markov property,

$$\sigma(\mathbf{o}_n) \underset{\sigma(\mathbf{s}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{o}_1, \ldots, \mathbf{o}_{n-1}),$$

and so, by Proposition A.6.1,

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid \mathbf{s}_n) = \mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_n)) \text{ a.s.},$$

for all $B \in \mathcal{O}$. Hence

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_n)) = \lambda\omega.\xi_n(\mathbf{s}_n(\omega))(B) \text{ a.s.}$$

that is,

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_n)) = \lambda\omega.(\lambda(h,a,s).\xi_n(s))((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_n)(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{O}$. In other words, $\lambda(h,a,s).\xi_n(s)$ is a regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{s}_n)$, for all $n \in \mathbb{N}$.

□

Proposition 2.3.2. *(Filter recurrence equations for state schemas) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(S, \mathcal{S})$ a state space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ a state process, $(\mu_n : H_n \to \mathcal{P}(S))_{n\in\mathbb{N}_0}$ the state schema, $(\tau_n : A \times S \to \mathcal{P}(S))_{n\in\mathbb{N}}$ the transition model, $(\xi_n : S \to \mathcal{P}(O))_{n\in\mathbb{N}}$ the observation model, υ_O a σ-finite measure on $\mathcal{O}$, and υ_S a σ-finite measure on $\mathcal{S}$. Suppose that the stochastic process $(\mathbf{a}, \mathbf{s}, \mathbf{o}) : \Omega \to A^{\mathbb{N}} \times S^{\mathbb{N}_0} \times O^{\mathbb{N}}$ is Markov. Suppose also that, for all $n \in \mathbb{N}_0$, there is a conditional density $\breve{\mu}_n : H_n \to \mathcal{D}(S)$ such that $\mu_n = \breve{\mu}_n \cdot \upsilon_S$ and that, for all $n \in \mathbb{N}$, there is a conditional density $\breve{\tau}_n : A \times S \to \mathcal{D}(S)$ such that $\tau_n = \breve{\tau}_n \cdot \upsilon_S$ and a conditional density $\breve{\xi}_n : S \to \mathcal{D}(O)$ such that $\xi_n = \breve{\xi}_n \cdot \upsilon_O$. Then the following hold.*

1. *For all $n \in \mathbb{N}_0$,*

$$\mu_{n+1} = \lambda(h,a,o).\lambda s.\breve{\xi}_{n+1}(s)(o) * \lambda(h,a,o).(\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))(h,a) \quad \mathcal{L}(\mathbf{h}_{n+1})\text{-a.e.}$$

2. *For all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,*

$$\mu_{n+1}(h_{n+1}) = \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * (\mu_n(h_n) \odot \lambda s.\tau_{n+1}(a_{n+1}, s)).$$

3. *For all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,*

$$\breve{\mu}_{n+1}(h_{n+1}) = \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * (\breve{\mu}_n(h_n) \odot \lambda s.\breve{\tau}_{n+1}(a_{n+1}, s)) \quad \upsilon_S\text{-a.e.}$$

Proof. 1. Each $\lambda(h,a).\breve{\mu}_n(h) : H_n \times A \to \mathcal{D}(S)$, $\lambda(h,a,s).\breve{\tau}_n(a,s) : H_{n-1} \times A \times S \to \mathcal{D}(S)$, and $\lambda(h,a,s).\breve{\xi}_n(s) : H_{n-1} \times A \times S \to \mathcal{D}(O)$ is a conditional density. Also

$$\begin{aligned}
&\lambda(h,a).\mu_n(h) = \lambda(h,a).\breve{\mu}_n(h) \cdot \upsilon_S, \text{ for all } n \in \mathbb{N}_0\\
&\lambda(h,a,s).\tau_{n+1}(a,s) = \lambda(h,a,s).\breve{\tau}_{n+1}(a,s) \cdot \upsilon_S, \text{ for all } n \in \mathbb{N}\\
&\lambda(h,a,s).\xi_n(s) = \lambda(h,a,s).\breve{\xi}_n(s) \cdot \upsilon_O, \text{ for all } n \in \mathbb{N}.
\end{aligned}$$

Towards the last of these, since $\xi_n = \breve{\xi}_n \cdot \upsilon_O$, it follows that

$$\begin{aligned}
&(\lambda(h,a,s).\breve{\xi}_n(s) \cdot \upsilon_O)(h,a,s)(B)\\
&= \int_O \mathbf{1}_B \, \lambda(h,a,s).\breve{\xi}_n(s)(h,a,s) \, d\upsilon_O\\
&= \int_O \mathbf{1}_B \, \breve{\xi}_n(s) \, d\upsilon_O\\
&= (\breve{\xi}_n \cdot \upsilon_O)(s)(B)\\
&= \xi_n(s)(B)\\
&= \lambda(h,a,s).\xi_n(s)(h,a,s)(B),
\end{aligned}$$

for all $(h,a,s) \in H_{n-1} \times A \times S$ and $B \in \mathcal{O}$. Hence $\lambda(h,a,s).\xi_n(s) = \lambda(h,a,s).\breve{\xi}_n(s) \cdot \upsilon_O$. The proofs of the first two are similar.

By Parts 1 and 2 of Proposition 2.3.1 and Proposition A.7.15,

$$\mathsf{P}(\mathbf{s}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.(\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))((\mathbf{h}_n, \mathbf{a}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $n \in \mathbb{N}_0$, and $C \in \mathcal{S}$. That is, for all $n \in \mathbb{N}_0$, $\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s) : H_n \times A \to \mathcal{P}(S)$ is a regular conditional distribution of $\mathbf{s}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

Also, since $\lambda(h,a).\mu_n(h) = \lambda(h,a).\breve{\mu}_n(h) \cdot \upsilon_S$ and $\lambda(h,a,s).\tau_{n+1}(a,s) = \lambda(h,a,s).\breve{\tau}_{n+1}(a,s) \cdot \upsilon_S$,

$$\begin{aligned}
&\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s)\\
&= (\lambda(h,a).\breve{\mu}_n(h) \cdot \upsilon_S) \odot (\lambda(h,a,s).\breve{\tau}_{n+1}(a,s) \cdot \upsilon_S)\\
&= (\lambda(h,a).\breve{\mu}_n(h) \odot \lambda(h,a,s).\breve{\tau}_{n+1}(a,s)) \cdot \upsilon_S. \qquad \text{[Proposition A.3.8]}
\end{aligned}$$

Hence $\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s) = (\lambda(h,a).\breve{\mu}_n(h) \odot \lambda(h,a,s).\breve{\tau}_{n+1}(a,s)) \cdot \upsilon_S$ and so $\lambda(h,a).\breve{\mu}_n(h) \odot \lambda(h,a,s).\breve{\tau}_{n+1}(a,s)$ is a regular conditional density of $\mathbf{s}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

By Part 3 of Proposition 2.3.1, $\lambda(h,a,s).\xi_{n+1}(s)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{s}_{n+1})$. Hence $\lambda(h,a,s).\breve{\xi}_n(s)$ is a regular conditional density.

Now consider the probability kernel

$$\begin{aligned}
&\lambda(h,a,o).\lambda s.\lambda(h,a,s).\breve{\xi}_{n+1}(s)(h,a,s)(o) \ast\\
&\qquad \lambda(h,a,o).(\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))(h,a) : H_{n+1} \to \mathcal{P}(S).
\end{aligned}$$

By Proposition A.12.8, this probability kernel is a regular conditional distribution of $\mathbf{s}_{n+1}$ given $\mathbf{h}_{n+1}$. Since μ_{n+1} is by definition a regular conditional distribution of $\mathbf{s}_{n+1}$ given

$\mathbf{h}_{n+1}$, the uniqueness part of Proposition A.5.16 shows that, $\mathcal{L}(\mathbf{h}_{n+1})$-almost everywhere,

$$\begin{aligned}
&\mu_{n+1}\\
&= \lambda(h,a,o).\lambda s.\lambda(h,a,s).\breve{\xi}_{n+1}(s)(h,a,s)(o) * \\
&\qquad\qquad \lambda(h,a,o).(\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))(h,a)\\
&= \lambda(h,a,o).\lambda s.\breve{\xi}_{n+1}(s)(o) * \lambda(h,a,o).(\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))(h,a).
\end{aligned}$$

2. For all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\mu_{n+1}(h_{n+1})\\
&= \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * (\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))(h_n, a_{n+1})\\
&= \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * (\mu_n(h_n) \odot \lambda s.\tau_{n+1}(a_{n+1}, s)).
\end{aligned}$$

3. For all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\breve{\mu}_{n+1}(h_{n+1}) \cdot \upsilon_S\\
&= \mu_{n+1}(h_{n+1})\\
&= \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * (\mu_n(h_n) \odot \lambda s.\tau_{n+1}(a_{n+1}, s))\\
&= \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * ((\breve{\mu}_n(h_n) \cdot \upsilon_S) \odot (\lambda s.\breve{\tau}_{n+1}(a_{n+1}, s) \cdot \upsilon_S))\\
&= \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * ((\breve{\mu}_n(h_n) \odot \lambda s.\breve{\tau}_{n+1}(a_{n+1}, s)) \cdot \upsilon_S) \qquad \text{[Proposition A.3.8]}\\
&= (\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * (\breve{\mu}_n(h_n) \odot \lambda s.\breve{\tau}_{n+1}(a_{n+1}, s))) \cdot \upsilon_S. \qquad \text{[Proposition A.3.10]}
\end{aligned}$$

The result now follows by Proposition A.2.11. □

In Part 1 of Proposition 2.3.2, the computation of

$$\lambda(h,a,o).(\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))(h,a)$$

is called the *transition update*, since it takes the current state schema and uses the transition model and current action to predict the new state schema. The projective product of $\lambda(h,a,o).\lambda s.\breve{\xi}_{n+1}(s)(o)$ with the output of the transition update is called the *observation update* because it takes the predicted state schema and uses the observation model to update the estimate of the state schema. Similarly, for Parts 2 and 3 Proposition 2.3.2. These are illustrated in Figure 2.8. Note that, given the initial state distribution $\mu_0(h_0)$, the recurrence equation in Part 2 of Proposition 2.3.2 enables the computation of $\mu_1(h_1)$, $\mu_2(h_2)$, $\mu_3(h_3)$, and so on.

Using the definition of the projective product, the recurrence equation for state distributions (in probability measure form) is more explicitly as follows: for all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\mu_{n+1}(h_{n+1})\\
&= \lambda B.\frac{\int_S \mathbf{1}_B \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ d(\mu_n(h_n) \odot \lambda s.\tau_{n+1}(a_{n+1}, s))}{\int_S \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ d(\mu_n(h_n) \odot \lambda s.\tau_{n+1}(a_{n+1}, s))}\\
&= \lambda B.\frac{\int_S \left(\lambda s'.\int_S \mathbf{1}_B \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ d\tau_{n+1}(a_{n+1}, s')\right)\ d\mu_n(h_n)}{\int_S \left(\lambda s'.\int_S \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ d\tau_{n+1}(a_{n+1}, s')\right)\ d\mu_n(h_n)}. \qquad \text{[Proposition A.7.8]}
\end{aligned}$$

$$\mu_{n+1} = \underbrace{\lambda(h,a,o).\lambda s.\breve{\xi}_{n+1}(s)(o)}_{observation\ update} * \underbrace{\lambda(h,a,o).(\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))(h,a)}_{transition\ update}$$

$$\mu_{n+1}(h_{n+1}) = \underbrace{\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})}_{observation\ update} * \underbrace{(\mu_n(h_n) \odot \lambda s.\tau_{n+1}(a_{n+1},s))}_{transition\ update}$$

$$\breve{\mu}_{n+1}(h_{n+1}) = \underbrace{\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})}_{observation\ update} * \underbrace{(\breve{\mu}_n(h_n) \odot \lambda s.\breve{\tau}_{n+1}(a_{n+1},s))}_{transition\ update}$$

Figure 2.8: Recurrence equations for state schemas and state distributions

Thus, for all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\mu_{n+1}(h_{n+1}) = \lambda B.\frac{\int_S \left(\lambda s'.\int_S \mathbf{1}_B \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ d\tau_{n+1}(a_{n+1},s')\right)\ d\mu_n(h_n)}{\int_S \left(\lambda s'.\int_S \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ d\tau_{n+1}(a_{n+1},s')\right)\ d\mu_n(h_n)}.$$

Similarly, the recurrence equation for state distributions in density form is more explicitly as follows: for all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$, υ_S-almost everywhere,

$$\begin{aligned}
&\breve{\mu}_{n+1}(h_{n+1}) \\
&= \frac{\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ (\breve{\mu}_n(h_n) \odot \lambda s.\breve{\tau}_{n+1}(a_{n+1},s))}{\int_S \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ (\breve{\mu}_n(h_n) \odot \lambda s.\breve{\tau}_{n+1}(a_{n+1},s))\ d\upsilon_S} \\
&= \frac{\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ \lambda s.\int_S \lambda s'.\breve{\tau}_{n+1}(a_{n+1},s')(s)\,\breve{\mu}_n(h_n)\ d\upsilon_S}{\int_S \left(\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ \lambda s.\int_S \lambda s'.\breve{\tau}_{n+1}(a_{n+1},s')(s)\,\breve{\mu}_n(h_n)\ d\upsilon_S\right)\ d\upsilon_S}.
\end{aligned}$$

Thus, for all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\breve{\mu}_{n+1}(h_{n+1}) = \frac{\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ \lambda s.\int_S \lambda s'.\breve{\tau}_{n+1}(a_{n+1},s')(s)\,\breve{\mu}_n(h_n)\ d\upsilon_S}{\int_S \left(\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ \lambda s.\int_S \lambda s'.\breve{\tau}_{n+1}(a_{n+1},s')(s)\,\breve{\mu}_n(h_n)\ d\upsilon_S\right)\ d\upsilon_S} \quad \upsilon_S\text{-a.e.}$$

The two preceding recurrence equations in explicit form are illustrated in Figure 2.9.

For readers familiar with the usual short derivation of the recurrence equations for state filtering in standard textbooks on AI and robotics ([140, Section 15.2.1], [141, Theorem 4.1], or [156, Section 2.4.3], for example), there may be some puzzlement at the

$$\mu_{n+1}(h_{n+1}) = \lambda B.\frac{\int_S \left(\lambda s'.\, \int_S \mathbf{1}_B \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\; d\tau_{n+1}(a_{n+1}, s')\right)\; d\mu_n(h_n)}{\int_S \left(\lambda s'.\, \int_S \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\; d\tau_{n+1}(a_{n+1}, s')\right)\; d\mu_n(h_n)}$$

$$\breve{\mu}_{n+1}(h_{n+1}) = \frac{\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\; \lambda s.\, \int_S \lambda s'.\breve{\tau}_{n+1}(a_{n+1}, s')(s)\, \breve{\mu}_n(h_n)\; d\upsilon_S}{\int_S \left(\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\; \lambda s.\, \int_S \lambda s'.\breve{\tau}_{n+1}(a_{n+1}, s')(s)\, \breve{\mu}_n(h_n)\; d\upsilon_S\right)\; d\upsilon_S}$$

Figure 2.9: Recurrence equations for state distributions in explicit form

complexity of the statement and proof of Proposition 2.3.2. Here is some insight about this matter. The usual derivation proceeds by starting with the state conditional density at time $n+1$. Bayes theorem is applied to this conditional density to obtain the product of a normalization constant and two conditional densities, the first of which is simplified by conditional independence assumptions to the observation model and the second is simplified, also using conditional independence assumptions, to an integral whose integrand is the product of the state conditional density at time n and the transition model. The resulting equation is the same as the second equation in Figure 2.9, taking into account the differing notation.

But there is a gap in this derivation. The three conditional densities in the numerator on the right hand side of the resulting equation certainly have the *same signature* as the conditional density forms of observation model, transition model, and state schema at time n, respectively, as can be inferred from the notation. The problem is that the derivation has not proved that they actually *are* the observation model, transition model, and state schema, respectively. The notations used in [140], [141], [156], and elsewhere suggest the intended meaning of the various conditional densities, but all that is actually ensured in the usual derivation is that the various conditional densities do have the correct intended signatures. However, specifying the signature does not uniquely specify the conditional density, of course, since distinct functions can have the same signature. To make the point with the more convenient probability kernels, the state schema component $\mu_n : H_n \to \mathcal{P}(S)$, for example, that appears on the right hand side of the recurrence equation is not just a probability kernel with the indicated signature, it is the (essentially unique) probability kernel with this signature that represents the conditional probability $\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n)$ via the condition $\mathsf{P}(\mathbf{s}_n^{-1}(C) \mid \mathbf{h}_n) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C)$ a.s., for all $C \in \mathcal{S}$.

It would be possible to complete the proof of the usual derivation by making sure that the necessary regular conditional density properties are preserved by the steps in the derivation. Here, instead, the result for conditional densities is obtained as a corollary of the result for probability kernels. And here the probability kernel proof proceeds in the opposite direction to the usual derivation, by showing that the right hand side of the equation is also a regular conditional distribution and therefore equal almost everywhere to the state schema at time $n+1$. This analysis shows why the concept of a regular conditional distribution is an essential ingredient of any theory of empirical beliefs.

Finally on this issue, the absence of various domain arguments in the transition and sensor models, and their corresponding conditional independence assumptions, actually complicates the proof of Proposition 2.3.2. For a proof not obscured by the absence of these domain arguments, see Proposition 4.1.2, which is arguably a more natural setting for the filter recurrence equations. From this point of view, Proposition 2.3.2 is an immediate corollary of Proposition 2.3.1 and Proposition 4.1.2.

Example 2.3.1. Consider a transition model $(\tau_n : A \times S \to \mathcal{P}(S))_{n\in\mathbb{N}}$ and $a \in A$ such that $\tau_n(a, s) = \delta_s$, for all $s \in S$ and $n \in \mathbb{N}$. Then, by Proposition A.2.7,

$$\mu_n(h_n) \odot \lambda s.\tau_{n+1}(a, s) = \mu_n(h_n),$$

for all $h_n \in H_n$ and $n \in \mathbb{N}_0$. In other words, for an action having such a transition model, there is no change to the state distribution during the transition update. That is, the action is a *no-op*.

Example 2.3.2. Suppose that the state is fixed over time. This situation can be captured by making the assumption that $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ is constant-valued almost surely, that is,

$$\mathsf{P}(\{\omega \mid \mathbf{s}_n(\omega) = \mathbf{s}_{n+1}(\omega), \text{ for all } n \in \mathbb{N}_0\}) = 1.$$

(The concept of 'constant-valued almost surely' is studied in more detail at the end of Section 4.1.) In this situation, one can assume that there are no actions nor is there a transition model to change the state and so the transition update part of filtering is skipped. Instead, for all $n \in \mathbb{N}_0$, since $\mu_n : H_n \to \mathcal{P}(S)$ is a regular conditional distribution of $\mathbf{s}_n$ given $\mathbf{h}_n$, the assumption that $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ is constant-valued almost surely shows that $\mu_n : H_n \to \mathcal{P}(S)$ is a regular conditional distribution of $\mathbf{s}_{n+1}$ given $\mathbf{h}_n$. A similar argument to the proof of Proposition 2.3.2 now shows that the second equation of Figure 2.9 reduces to

$$\breve{\mu}_{n+1}(h_{n+1}) = \frac{\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ \breve{\mu}_n(h_n)}{\int_S \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1})\ \breve{\mu}_n(h_n)\ d\upsilon_S},$$

which is Bayes theorem. Thus filtering is a generalization of Bayesian inference.

Example 2.3.3. Suppose that $(\xi_n : S \to \mathcal{P}(O))_{n\in\mathbb{N}}$ is an observation model such that, for all $n \in \mathbb{N}$ and $o \in O$, $\lambda s.\breve{\xi}_n(s)(o) : S \to \mathbb{R}$ is a constant function, which may vary with n and o. This is the case, for example, if O is a singleton set. Then

$$\begin{aligned} &\mu_{n+1}(h_{n+1}) \\ &= \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * (\mu_n(h_n) \odot \lambda s.\tau_{n+1}(a_{n+1}, s)) \\ &= \mu_n(h_n) \odot \lambda s.\tau_{n+1}(a_{n+1}, s). \end{aligned}$$

Thus ξ is an observation model that provides no information for the observation update, which therefore does not change the distribution produced by the transition update. Such an observation model is called *non-informative*.

Example 2.3.4. This example considers the case where the state space S in Proposition 2.3.2 is finite, which is commonly known as the (finite state) hidden Markov model case. As would be done in a practical application, the example concentrates on the recurrence equation for state distributions in density form:

$$\breve{\mu}_{n+1}(h_{n+1}) = \lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) * (\breve{\mu}_n(h_n) \odot \lambda s.\breve{\tau}_{n+1}(a_{n+1}, s)).$$

The underlying measure space is $(S, 2^S, c)$, where c is counting measure, and the state distribution is categorical. Suppose that $S = \{s_1, \ldots, s_m\}$. Then

$$\begin{aligned}
&\breve{\mu}_{n+1}(h_{n+1})(s_i) \\
&= \frac{\breve{\xi}_{n+1}(s_i)(o_{n+1}) \ \int_S \lambda s'.\breve{\tau}_{n+1}(a_{n+1}, s')(s_i)\, \breve{\mu}_n(h_n)\, dc}{\int_S \left(\lambda s.\breve{\xi}_{n+1}(s)(o_{n+1}) \ \lambda s.\ \int_S \lambda s'.\breve{\tau}_{n+1}(a_{n+1}, s')(s)\, \breve{\mu}_n(h_n)\, dc\right)\ dc} \qquad \text{[Figure 2.9]} \\
&= \frac{\breve{\xi}_{n+1}(s_i)(o_{n+1}) \sum_{k=1}^m \breve{\tau}_{n+1}(a_{n+1}, s_k)(s_i)\ \breve{\mu}_n(h_n)(s_k)}{\sum_{j=1}^m (\breve{\xi}_{n+1}(s_j)(o_{n+1}) \sum_{k=1}^m \breve{\tau}_{n+1}(a_{n+1}, s_k)(s_j)\ \breve{\mu}_n(h_n)(s_k))}, \qquad \text{[Example A.2.6]}
\end{aligned}$$

for $i = 1, \ldots, m$. Consequently, $\breve{\mu}_{n+1}(h_{n+1})$ is a tractable expression (provided that m is not too large). Note that this result depends only on the finiteness of S; there are no restrictions on A or O.

Example 2.3.5. Continuing Example 2.3.4, let the observation space O be S and each component $\xi_n : S \to \mathcal{P}(S)$ of the observation model ξ be defined by $\xi_n(s) = \delta_s$, for all $s \in S$. Then $\breve{\xi}_n : S \to \mathcal{D}(S)$ is defined by

$$\breve{\xi}_n(s)(s') = \begin{cases} 1 & \text{if } s' = s \\ 0 & \text{otherwise,} \end{cases}$$

for all $s, s' \in S$. Let $o_{n+1} = \overline{s} \in S$. Then, for all $i = 1, \ldots, m$,

$$\begin{aligned}
&\breve{\mu}_{n+1}(h_{n+1})(s_i) \\
&= \frac{\breve{\xi}_{n+1}(s_i)(\overline{s}) \sum_{k=1}^m \breve{\tau}_{n+1}(a_{n+1}, s_k)(s_i)\ \breve{\mu}_n(h_n)(s_k)}{\sum_{j=1}^m (\breve{\xi}_{n+1}(s_j)(\overline{s}) \sum_{k=1}^m \breve{\tau}_{n+1}(a_{n+1}, s_k)(s_j)\ \breve{\mu}_n(h_n)(s_k))} \\
&= \begin{cases} 1 & \text{if } s_i = \overline{s} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Thus, whatever distribution is produced in the transition update, the observation update with this observation model picks out the observation $\overline{s}$ as the next state. This is the fully observable case of Markov decision processes. Such an observation model is called *perfect*.

Example 2.3.6. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(S, \mathcal{S})$ a state space, $(O, \mathcal{O})$ and $(O', \mathcal{O}')$ observation spaces, $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ a state process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, and $\xi \triangleq (\xi_n : S \to \mathcal{P}(O))_{n \in \mathbb{N}}$ an observation model. Suppose that $p : O \to O'$ is a measurable function. Then, according to Proposition A.10.2,

$$\lambda s.(\xi_n(s) \circ p^{-1}) : S \to \mathcal{P}(O')$$

is a regular conditional distribution of $p \circ \mathbf{o}_n$ given $\mathbf{s}_n$, for all $n \in \mathbb{N}$. Hence $(\lambda s.(\xi_n(s) \circ p^{-1}) : S \to \mathcal{P}(O'))_{n\in\mathbb{N}}$ is an observation model, called the *quotient observation model* (from the observation model ξ via the mapping p).

In an application, if O is a complicated space, it may be difficult to learn or specify ξ. In such cases, a suitable mapping p can be used to reduce the observation space to a simpler space O', which could even be as simple as $\mathbb{B}^m$ or $\mathbb{R}^m$. This may lead to an observation model $(\lambda s.(\xi_n(s) \circ p^{-1}) : S \to \mathcal{P}(O'))_{n\in\mathbb{N}}$ that is easier to learn or specify. However, there is generally a cost involved in passing to the quotient observation model as the new observation space O' is less 'precise' that the original observation space, being a set of equivalence classes of it. Thus the quotient observation model generally provides less information for the observation update than the original observation model does. In the extreme case when O' is a singleton set, the corresponding quotient observation model does not change at all the state distribution that is obtained from the transition update. (See Example 2.3.3.)

Some remarks on the observation space may be useful. There are no theoretical restrictions on the choice of observation space (apart from being a standard Borel space). However, there is a practical restriction: the observation space must be observable, of course. Also some observation spaces are better than others. A good observation space is one such that the corresponding observation model is able to sharply distinguish different state spaces depending on the observation that is observed; in this case, the observation update produces a tighter distribution around the actual state. Example 2.3.6 shows how moving to a quotient observation model may be useful, although this generally involves a trade-off. But note carefully that moving to another (different) observation space means that the definition of history changes and hence the definition of state schema also changes. However, once a state distribution has been obtained from a state schema by instantiating with the (thus modified) history, this change becomes unimportant.

Example 2.3.7. This example considers the case where the state distribution is Gaussian. In this setting, the state space S is $\mathbb{R}^m$, for some $m \in \mathbb{N}$. (For example, a state could be the pose of a robot in a planar environment, which is a 3-dimensional vector consisting of the two cartesian coordinates and the angular orientation.) The action space A is $\mathbb{R}^k$, for some $k \in \mathbb{N}$. (For example, an action for a robot in a planar environment could be a 2-dimensional vector specifying the translational and rotational velocities that are applied to the robot.) The observation space O is $\mathbb{R}^p$, for some $p \in \mathbb{N}$. (For example, components of an observation could be range values from a range finder, brightness values from a camera, and so on.) The filter described here is the Kalman filter [84].

Let $(\nu_n : H_n \to \mathcal{P}(\mathbb{R}^m))_{n\in\mathbb{N}_0}$ be the relevant state schema. Suppose that the initial distribution $\breve{\nu}_0(h_0)$ for the state is a Gaussian density with mean μ_0 and covariance Σ_0; hence

$$\breve{\nu}_0(h_0) = \lambda s.\frac{1}{(2\pi)^{m/2}}\frac{1}{|\Sigma_0|^{1/2}}\exp\left\{-\frac{1}{2}(s-\mu_0)^T\Sigma_0^{-1}(s-\mu_0)\right\}.$$

Next the transition model is specified. For all $n \in \mathbb{N}$, let A_n be an $m \times m$ matrix of real values and B_n an $m \times k$ matrix of real values. For all $n \in \mathbb{N}$, let R_n be an $m \times m$ covariance matrix. The transition model is a linear Gaussian: for all $n \in \mathbb{N}$, $\breve{\tau}_n : A \times S \to \mathcal{D}(S)$ is

defined by

$$\breve{\tau}_n(a,s) = \lambda s'.\frac{1}{(2\pi)^{m/2}}\frac{1}{|R_n|^{1/2}}\exp\left\{-\frac{1}{2}(s'-A_ns-B_na)^TR_n^{-1}(s'-A_ns-B_na)\right\},$$

for all $s \in S$ and $a \in A$.

Now the observation model is defined. For all $n \in \mathbb{N}$, let C_n be an $p \times m$ matrix of real values and Q_n a $p \times p$ covariance matrix. The observation model is a linear Gaussian: for all $n \in \mathbb{N}$, $\breve{\xi}_n : S \to \mathcal{D}(O)$ is defined by

$$\breve{\xi}_n(s) = \lambda o.\frac{1}{(2\pi)^{k/2}}\frac{1}{|Q_n|^{1/2}}\exp\left\{-\frac{1}{2}(o-C_ns)^TQ_n^{-1}(o-C_ns)\right\},$$

for all $s \in S$.

Assume now that the state distribution $\breve{\nu}_n(h_n)$ at time step n is a Gaussian density with mean μ_n and covariance Σ_n. Then the distribution $\breve{\nu}_{n+1}(h_{n+1})$ at time step $n+1$ is a Gaussian distribution with mean μ_{n+1} and covariance Σ_{n+1}, where

$$\begin{aligned}
\mu_{n+1} &= \overline{\mu}_{n+1} + K_{n+1}(o_{n+1} - C_{n+1}\overline{\mu}_{n+1})\\
\Sigma_{n+1} &= (I - K_{n+1}C_{n+1})\overline{\Sigma}_{n+1}\\
K_{n+1} &= \overline{\Sigma}_{n+1}C_{n+1}^T(C_{n+1}\overline{\Sigma}_{n+1}C_{n+1}^T + Q_{n+1})^{-1}\\
\overline{\mu}_{n+1} &= A_{n+1}\mu_n + B_{n+1}a_{n+1}\\
\overline{\Sigma}_{n+1} &= A_{n+1}\Sigma_nA_{n+1}^T + R_{n+1}.
\end{aligned}$$

For the proof of this, see [156, Section 3.2.4]. Consequently, for all $n \in \mathbb{N}$, $\breve{\nu}_{n+1}(h_{n+1})$ is a Gaussian density and a tractable expression.

Example 2.3.8. This example considers filtering distributions on the state $\mathbb{B}^m$, for $m \in \mathbb{N}$. For this setting, $\breve{\mu}_n : H_n \to \mathcal{D}(\mathbb{B}^m)$, for all $n \in \mathbb{N}_0$, and $\breve{\tau}_n : A \times \mathbb{B}^m \to \mathcal{D}(\mathbb{B}^m)$, and $\breve{\xi}_n : \mathbb{B}^m \to \mathcal{D}(O)$, for all $n \in \mathbb{N}$.

Let $\breve{\mu}_0(h_0) : \mathcal{D}(\mathbb{B}^m)$ be the density defined in Example A.3.6 that requires m parameters for its definition. This is the initial state distribution.

Let $g : \mathcal{D}(\mathbb{B}^m)$ be the density also defined in Example A.3.6 that requires $2^m - 1$ parameters for its definition. Suppose that, for all $n \in \mathbb{N}$ and for some action $a \in A$, $\lambda s.\breve{\tau}_n(a,s) : \mathbb{B}^m \to \mathcal{D}(\mathbb{B}^m)$ is defined by

$$\breve{\tau}_n(a,s) = g,$$

for all $s \in \mathbb{B}^m$.

Let $\breve{\xi}_n$ be a non-informative observation model, for all $n \in \mathbb{N}$, as explained in Example 2.3.3. Then

$$\begin{aligned}
&\breve{\mu}_1(h_1)\\
&= \lambda s.\sum_{s' \in \mathbb{B}^m}\breve{\tau}_1(a,s')(s)\ \breve{\mu}_0(h_0)(s')\\
&= g\sum_{s' \in \mathbb{B}^m}\breve{\mu}_0(h_n)(s')\\
&= g.
\end{aligned}$$

In fact, every time the action is a, $\breve{\mu}_n(h_n) = g$. Thus the transition update changes the initial state distribution needing m parameters to a state distribution needing $2^m - 1$ parameters, which can be intractable to even store for values of m likely to occur in practice. Note how the transition update has added many conditional dependencies to $\breve{\mu}_1(h_1)$ that did not occur in $\breve{\mu}_0(h_0)$ and this is the fundamental reason why $\breve{\mu}_1(h_1)$ becomes intractable.

The transition model for this example is thoroughly implausible for any practical application. Nevertheless, the example does show that filtering can run into problems of intractability. In fact, this phenomenom does occur in practical problems and is known as *entanglement*. It is an impediment to filtering in the case of structured state spaces.

Here is a version of Bayes theorem (Proposition A.7.14) specialized for state schemas.

Notation. For $A \subseteq X_1 \times X_2$, put $A^* = \{(x_2, x_1) \mid (x_1, x_2) \in A\}$.

Proposition 2.3.3. *(Bayes theorem for state schemas) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(S, \mathcal{S})$ a state space, $(O, \mathcal{O})$ an observation space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ a state process, and $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process. Let Ξ be the environment for $\mathbf{a}$ and $\mathbf{o}$, μ the state schema, ξ the observation model, and τ the transition model. Suppose that the stochastic process $(\mathbf{a}, \mathbf{s}, \mathbf{o}) : \Omega \to A^{\mathbb{N}} \times S^{\mathbb{N}_0} \times O^{\mathbb{N}}$ is Markov. Then, for all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost everywhere,*

$$\begin{aligned}&\lambda(h,a).(\Xi_{n+1} \otimes \mu_{n+1})(h,a)(E^*) = \\ &\quad \lambda(h,a).((\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s)) \otimes \lambda(h,a,s).\xi_{n+1}(s))(h,a)(E),\end{aligned}$$

for all $E \in \mathcal{S} \otimes \mathcal{O}$.

Proof. As shown in the proof of Proposition 2.3.2,

$$\mathsf{P}(\mathbf{s}_{n+1}^{-1}(D) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.(\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s))((\mathbf{h}_n, \mathbf{a}_{n+1})(\omega))(D) \text{ a.s.},$$

for all $D \in \mathcal{S}$. Thus $\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s) : H_n \times A \to \mathcal{P}(S)$ is a regular conditional distribution of $\mathbf{s}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

By Part 3 of Proposition 2.3.1,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{s}_{n+1})) = \lambda\omega.\lambda(h,a,s).\xi_{n+1}(s)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{s}_{n+1})(\omega))(B),$$

for all $B \in \mathcal{O}$. Thus $\lambda(h,a,s).\xi_{n+1}(s)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{s}_{n+1})$.

By Proposition A.7.12, for all $n \in \mathbb{N}_0$ and $E \in \mathcal{S} \otimes \mathcal{O}$,

$$\begin{aligned}&\mathsf{P}((\mathbf{s}_{n+1}, \mathbf{o}_{n+1})^{-1}(E) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \\ &\lambda\omega.((\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s)) \otimes \lambda(h,a,s).\xi_{n+1}(s))((\mathbf{h}_n, \mathbf{a}_{n+1})(\omega))(E) \text{ a.s.}\end{aligned}$$

Now consider $\Xi_{n+1} \otimes \mu_{n+1}$. By Proposition A.7.12, for all $n \in \mathbb{N}_0$ and $E \in \mathcal{S} \otimes \mathcal{O}$,

$$\mathsf{P}((\mathbf{o}_{n+1}, \mathbf{s}_{n+1})^{-1}(E^*) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.(\Xi_{n+1} \otimes \mu_{n+1})((\mathbf{h}_n, \mathbf{a}_{n+1})(\omega))(E^*) \text{ a.s.}$$

Thus, by the uniqueness part of Proposition A.5.16, it follows that, for all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost everywhere,

$$\lambda(h,a).(\Xi_{n+1} \otimes \mu_{n+1})(h,a)(E^*) = \\ \lambda(h,a).((\lambda(h,a).\mu_n(h) \odot \lambda(h,a,s).\tau_{n+1}(a,s)) \otimes \lambda(h,a,s).\xi_{n+1}(s))(h,a)(E),$$

for all $E \in \mathcal{S} \otimes \mathcal{O}$. □

Proposition 2.3.2 provides the basis for a doxastically rational agent that employs the state distribution as its (primary) empirical belief for deciding how to act. At each time step, the agent uses the known effect of its action and the consequent observation to determine the new state distribution from the previous one. Such an agent would also be instrumentally rational if, for example, it used a utility function on states to determine which action maximized its expected utility and then acted accordingly.

To set the scene for the following development in this book, some discussion of Proposition 2.3.2 is appropriate. For this, the discussion is mostly about the filter recurrence equations for state distributions, the second and third equations in Figure 2.8. If $\mu_n(h_n)$ is the current state distribution, then the filter recurrence equation precisely gives the next state distribution $\mu_{n+1}(h_{n+1})$. However, in many applications, filtering causes the syntactic complexity of updated state distributions to increase rapidly making dealing directly with the mathematical form of the updated empirical beliefs impractical after only a small number of time steps. There are a few important cases for which this problem can be avoided. If the state distribution is a categorical distribution on a finite space (the hidden Markov model case), then Example 2.3.4 shows that the updated state distribution is a tractable expression, which avoids the problem. If the state distribution is a Gaussian, and the transition and observation models are linear Gaussian (the Kalman filter case), then Example 2.3.7 shows that the updated state distribution is another Gaussian, which again avoids the problem.

However, for the cases of interest in this book, states are often much more complex than those in Examples 2.3.4 and 2.3.7. For example, the setting of this section is general enough to include dynamic Bayesian networks as a special case. The approach to filtering adopted here is intended to be able to deal with state (and observation) spaces that have complex structure. First, an argument is given for why this is necessary. Then, the approach to deal with the structure is outlined.

A common way of dealing with states is to map them into features. Here is how this is done. Consider the state process $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$, from which one obtains the random variable $\mathbf{s}_n : \Omega \to S$, for all $n \in \mathbb{N}_0$. Elements in S are assumed to have the detail required to provide a model of states that accurately reflects all the information about the environment that will be needed for the application. However, instead of dealing with distributions on S, one can proceed as follows.

A *feature* is a property or characteristic of an individual that is identified for some purpose, often for that of learning. Formally, a feature is an element of a 'simple' space such as the reals $\mathbb{R}$, the non-negative integers $\mathbb{N}_0$, the booleans $\mathbb{B}$, or a (small) finite set. Each such space is called a *feature space*. Typical distributions on feature spaces are, for example, Gaussians, Poissons, or categorical distributions. A (measurable) function p, called a *feature mapping*, is defined from S into a product space $\prod_{i=1}^{k} X_i$, where each X_i

is a feature space. Let S' denote the space $\prod_{i=1}^{k} X_i$. The elements of S' are called *feature vectors* and distributions on S' are convenient to work with being products of Gaussian, Poisson, or categorical distributions, or a combination of these. This approach normally requires the designer to specify the function p, which involves selecting suitable features. Utility functions on S' can be defined to assist the selection of actions. The space S' thus becomes the *de facto* state space.

For some applications, such an approach is fine. The designer *is* able to define an effective p and the utility function defined on S' is sufficiently accurate so that good choices of actions can be made. However, in general, there may be some significant problems with the approach. The first is that it may be too much to expect the designer to be able to find a sufficiently good function p. The second is that the utility function should actually be defined on S, not S'. In fact, the utility function $U : S' \to \mathbb{R}$ does induce a utility $U \circ p : S \to \mathbb{R}$ on S. But there are, in general, many more possible utility functions on S that cannot be factored through S' in this way. In effect, the utility functions on S that are obtained by factoring through S' must be constant on the equivalence classes induced by p and, unless the partition is just right, the corresponding utility function may not be effective enough.

The current state distribution should assist the agent to communicate usefully with other agents and humans. For this purpose, it is obvious that the more detailed the state, the more information is available to the agent to undertake this communication. More precisely, it is the *distribution* on the state space that the agent uses for this purpose. Now S' contains less information about the environment than S by its construction: each element of S' corresponds to an equivalence class of states in S. If the partition on S is too coarse, the distribution on S' may not provide enough information for the agent to effectively communicate. (Note that, if μ is a probability measure on S, then $\mu \circ p^{-1}$ is a probability measure on S', by Proposition A.2.13; the distributions on S and S' can be expected to be related in this way.)

Thus the crucial issues are: how fine does the partition on S induced by p have to be to allow an agent to act effectively, and how difficult is it for the designer to find the right partition? Ultimately this is a design decision, but the more detail available in the states in S' about the environment the better. In fact, ideally, for this reason, it is best to deal directly with states in S and avoid having to find p altogether. Thus the technical issue of dealing directly with distributions on arbitrarily structured spaces arises.

In general, a state space can be a product space, a sum space, or a quotient space at the top level with embedded product, sum, or quotient spaces inside. The product spaces can even be *infinite* dimensional in the case that there are sets or multisets present in the state. So a state can be a highly structured object. Instead of dealing with a distribution on the state as a whole, the obvious approach is to deconstruct the state distribution into simpler probabilistic components and deal with them separately. In all cases, the simpler probabilistic components are probability kernels that can be recombined back into the state distribution. In terms of empirical beliefs, the state distribution, an empirical belief in the form of a probability measure, can be deconstructed into a number of (more or less arbitrary) empirical beliefs that are probability kernels. Then the filter recurrence equations for state distributions will need to be extended to (arbitrary) empirical beliefs.

In the light of the above analysis and the research program it suggests, here is a summary of what is coming in later chapters:

1. Chapter 3 considers how to systematically construct and deconstruct empirical beliefs, and how to use empirical beliefs for belief representation purposes.
2. Chapter 4 considers the extension of the filter recurrence equations of Proposition 2.3.2 to arbitrary empirical beliefs.
3. Chapter 5 considers how an agent can reason with its empirical beliefs to help choose its actions.

Bibliographical Notes

The general setting for agents and environments employed here is similar to that of [78].

Another common approach to agents is to employ Markov decision processes [140, Section 17.1]. These are defined by the assumptions that the environment can be at any time in one of a set of states, each observation is a state, the next observation emitted by the environment depends only on the current state and the last action of the agent, and, finally, it is sufficient for the agent to select actions that depend only on the current state. In this case, the states are fully observable by the agent. More generally, the states may be hidden and only partially observable by the agent through observations. In this case, the agent usually has, or can learn, an observation model that gives a distribution on observations that would be emitted given that the environment is in a particular state. This weaker assumption defines partially observable Markov decision processes [140, Section 17.4]. This latter setting is that of Section 2, although the terminology and formalization here are different.

The setting of Section 2 also includes as a special case that of dynamic Bayesian networks, which are discussed, for example, in [140, Section 15.5] and [88, Section 6.2.2]: just let S be a product space and assume the definitions of the exogenous transition model and the observation model take advantage of the sparseness of the underlying Bayesian network.

For the simpler setting of (elementary) conditional probabilities, Proposition 2.3.2 is discussed, for example, in [140, Section 15.2.1] and [156, Section 2.4.3]. In that setting, the (analogue of the) condition $\sigma(\mathbf{o}_n) \perp\!\!\!\perp_{\sigma(\mathbf{s}_n)} \sigma(\mathbf{a}_1, \dots \mathbf{a}_n, \mathbf{o}_1, \dots, \mathbf{o}_{n-1})$, for all $n \in \mathbb{N}$, in Proposition 2.3.2 is called the sensor Markov assumption in [140].

The entanglement problem is discussed in [88, Section 15.2.4], [113, Section 17.6.7], and [140, Section 15.5.2].

Exercises

2.1 Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, and $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process. Prove that the following are equivalent.

1. $\sigma(\mathbf{a}_n) \underset{\sigma(\mathbf{a}_1,\dots,\mathbf{a}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{o}_1, \dots, \mathbf{o}_{n-1})$,

for all $n \in \mathbb{N}$.

2. $\sigma(\mathbf{o}_1, \dots, \mathbf{o}_n) \underset{\sigma(\mathbf{a}_1,\dots,\mathbf{a}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{a}_{n+1}, \dots, \mathbf{a}_{n+k})$,

for all $n \in \mathbb{N}$ and $k \in \mathbb{N}$.

3. $\sigma(\mathbf{o}_n) \perp\!\!\!\perp_{\sigma(\mathbf{a}_1,\mathbf{o}_1,\ldots,\mathbf{a}_{n-1},\mathbf{o}_{n-1},\mathbf{a}_n)} \sigma(\mathbf{a}_{n+1},\ldots,\mathbf{a}_{n+k})$,

for all $n \in \mathbb{N}$ and $k \in \mathbb{N}$.

4. $\sigma(\mathbf{o}_n) \perp\!\!\!\perp_{\sigma(\mathbf{a}_1,\mathbf{o}_1,\ldots,\mathbf{a}_{n-1},\mathbf{o}_{n-1},\mathbf{a}_n)} \sigma(\mathbf{a}_{n+1},\mathbf{a}_{n+2},\ldots)$,

for all $n \in \mathbb{N}$.

2.2 An agent Λ is *deterministic* if Λ_n is a Dirac kernel, for all $n \in \mathbb{N}$. Thus an agent Λ is deterministic precisely when $\Lambda_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1})$ is a Dirac measure on A, for all $(a_1, o_1, \ldots, a_{n-1}, o_{n-1}) \in H_{n-1}$ and for all $n \in \mathbb{N}$. In other words, a deterministic agent is a sequence $(\Lambda_n)_{n\in\mathbb{N}}$, where

$$\Lambda_n : H_{n-1} \to \Delta(A)$$

is a measurable function, for all $n \in \mathbb{N}$.

1. Let $(\Lambda_n)_{n\in\mathbb{N}}$ be a deterministic agent. Prove that there exists a unique sequence $(\bar{\Lambda}_n)_{n\in\mathbb{N}}$, where $\bar{\Lambda}_n : O^{n-1} \to A$ is a measurable function, for all $n \in \mathbb{N}$, such that

$$\bar{\Lambda}_n(o_1, \ldots, o_{n-1}) = a_n, \text{ where } \begin{aligned} &\Lambda_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}) = \delta_{a_n}, \\ &\qquad\vdots \\ &\Lambda_2(a_1, o_1) = \delta_{a_2}, \\ &\Lambda_1() = \delta_{a_1}, \end{aligned}$$

   for all $n \in \mathbb{N}$ and for all $(o_1, \ldots, o_{n-1}) \in O^{n-1}$.

2. Let $(\bar{\Lambda}_n)_{n\in\mathbb{N}}$ be a sequence such that $\bar{\Lambda}_n : O^{n-1} \to A$ is a measurable function, for all $n \in \mathbb{N}$. Prove that there exists a deterministic agent $(\Lambda_n)_{n\in\mathbb{N}}$ such that

$$\Lambda_n(\bar{\Lambda}_1(), o_1, \ldots, \bar{\Lambda}_n(o_1, \ldots, o_{n-2}), o_{n-1}) = \delta_{a_n}, \text{ where } \bar{\Lambda}_n(o_1, \ldots, o_{n-1}) = a_n,$$

   for all $(o_1, \ldots, o_{n-1}) \in O^{n-1}$ and for all $n \in \mathbb{N}$.

It follows that a deterministic agent can be identified with a sequence $(\Lambda_n)_{n\in\mathbb{N}}$, where $\Lambda_n : O^{n-1} \to A$ is a measurable function, for all $n \in \mathbb{N}$.

2.3 An environment Ξ is *deterministic* if Ξ_n is a Dirac kernel, for all $n \in \mathbb{N}$. Thus an environment Ξ is deterministic precisely when $\Xi_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}, a_n)$ is a Dirac measure on O, for all $(a_1, o_1, \ldots, a_{n-1}, o_{n-1}, a_n) \in H_{n-1} \times A$ and for all $n \in \mathbb{N}$. In other words, a deterministic environment is a sequence $(\Xi_n)_{n\in\mathbb{N}}$, where

$$\Xi_n : H_{n-1} \times A \to \Delta(O)$$

is a measurable function, for all $n \in \mathbb{N}$.

1. Let $(\Xi_n)_{n\in\mathbb{N}}$ be a deterministic environment. Prove that there exists a unique sequence $(\bar{\Xi}_n)_{n\in\mathbb{N}}$, where $\bar{\Xi}_n : A^n \to O$ is a measurable function, for all $n \in \mathbb{N}$, such that

$$\begin{aligned}\bar{\Xi}_n(a_1,\ldots,a_n) = o_n, \text{ where } \Xi_n(a_1,o_1,\ldots,a_{n-1},o_{n-1},a_n) &= \delta_{o_n},\\ &\vdots\\ \Xi_2(a_1,o_1,a_2) &= \delta_{o_2},\\ \Xi_1(a_1) &= \delta_{o_1},\end{aligned}$$

   for all $n \in \mathbb{N}$ and for all $(a_1,\ldots,a_n) \in A^n$.

2. Let $(\bar{\Xi}_n)_{n\in\mathbb{N}}$ be a sequence such that $\bar{\Xi}_n : A^n \to O$ is a measurable function, for all $n \in \mathbb{N}$. Prove that there exists a deterministic environment $(\Xi_n)_{n\in\mathbb{N}}$ such that

$$\Xi_n(a_1,\bar{\Xi}_1(a_1),\ldots,a_{n-1},\bar{\Xi}_{n-1}(a_1,\ldots,a_{n-1}),a_n) = \delta_{o_n}, \text{ where } \bar{\Xi}_n(a_1,\ldots,a_n) = o_n,$$

   for all $(a_1,\ldots,a_n) \in A^n$ and for all $n \in \mathbb{N}$.

It follows that a deterministic environment can be identified with a sequence $(\Xi_n)_{n\in\mathbb{N}}$, where $\Xi_n : A^n \to O$ is a measurable function, for all $n \in \mathbb{N}$.

2.4 Compare and contrast the statement and proof of Proposition 2.3.2 with the corresponding account in [156, Section 2.4.3] for the case of (elementary) conditional probabilities. Is the greater technical sophistication of Proposition 2.3.2 justified?

2.5 Prove the correctness of the recurrence equations for μ_n and Σ_n given in Example 2.3.7.

Chapter 3

Structure of Empirical Beliefs

THIS chapter begins by introducing the two main concepts of this book, schemas and empirical beliefs. Empirical beliefs are obtained from schemas by instantiating a schema with the current history. Schemas, in contrast to empirical beliefs, are regular conditional distributions, a fact that is used to prove many of their properties. The construction of more complex schemas from simpler ones and the deconstruction of complex schemas into simpler ones are studied. The connections between the mathematical concepts of product and sum, and typical data types such as tuples, lists, strings, sequences, sets, multisets, and graphs are established. The chapter contains various examples to illustrate the representation issues.

3.1 Schemas and Empirical Beliefs

To assist an agent in deciding which actions to perform in order to achieve its goal(s), the agent has a belief base that consists of certain function definitions, embodying information about, for example, the environment, its location, and beliefs of other agents. Included in the belief base are empirical beliefs, those beliefs acquired from observations. The discussion of empirical beliefs starts with the prior concept of a schema.

In the following definitions, the agent-environment setting from Chapter 2 is employed. Thus there is an action space $(A, \mathcal{A})$, an observation space $(O, \mathcal{O})$, an action process $\mathbf{a} : \Omega \to A^{\mathbb{N}}$, and an observation process $\mathbf{o} : \Omega \to O^{\mathbb{N}}$. For all $n \in \mathbb{N}_0$, $H_n \triangleq A \times O \times \cdots \times A \times O$, where there are n occurrences of A and n occurrences of O, and $\mathbf{h}_n \triangleq (\mathbf{a}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{o}_n) : \Omega \to H_n$. If $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$, then, for all $n \in \mathbb{N}_0$, $\mathbf{x}_n : \Omega \to X$ is defined by $\mathbf{x}_n(\omega) = \mathbf{x}(\omega)(n)$, for all $\omega \in \Omega$. Similarly, for $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ and $\mathbf{y}_n : \Omega \to Y$.

Now comes the first of two main concepts of this book.

Definition 3.1.1. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ measurable spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes. A *schema* (*for* $\mathbf{y}$ *given* $\mathbf{x}$) is a sequence $\mu \triangleq (\mu_n)_{n \in \mathbb{N}_0}$, where

$$\mu_n : H_n \times X \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{x}_n)$, for all $n \in \mathbb{N}_0$.

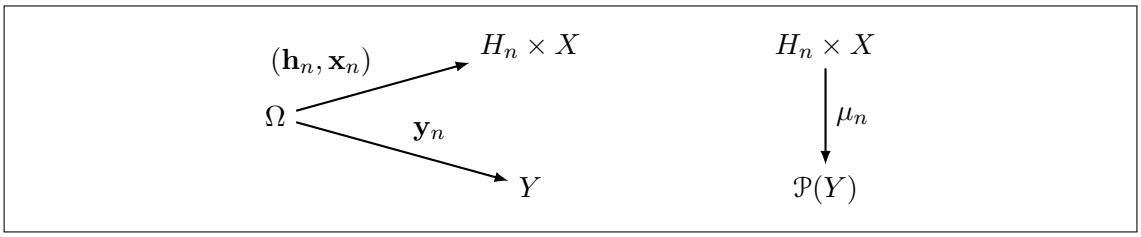

Figure 3.1: A component of a schema

In other words, for all $n \in \mathbb{N}_0$, μ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{y}_n^{-1}(B) \mid (\mathbf{h}_n, \mathbf{x}_n)) = \lambda\omega.\mu_n((\mathbf{h}_n, \mathbf{x}_n)(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{Y}$. Each μ_n is referred to as a *component* of the schema.

Note that, for all $n \in \mathbb{N}_0$, $\mu_n \circ (\mathbf{h}_n, \mathbf{x}_n) : \Omega \to \mathcal{P}(Y)$ is a random probability measure. Similarly, $(\mu_n \circ (\mathbf{h}_n, \mathbf{x}_n))_{n\in\mathbb{N}_0} : \Omega \to \mathcal{P}(Y)^{\mathbb{N}_0}$ is a stochastic process.

By omitting X (or, equivalently, by letting X be a distinguished singleton set), the preceding definition includes the case when each schema component has the form $\mu_n : H_n \to \mathcal{P}(Y)$. In this case, the schema is said to be *for* $\mathbf{y}$.

Example 3.1.1. A prototypical example of a schema is a state schema from Definition 2.3.4 having components with signature

$$\mu_n : H_n \to \mathcal{P}(S),$$

where S is a state space and $\mathbf{s} : \Omega \to S^{\mathbb{N}_0}$ is a state process. This is a schema for $\mathbf{s}$.

If Y is a standard Borel space, then schemas are guaranteed to exist.

Proposition 3.1.1. *(Existence of schemas) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ a measurable space, $(Y, \mathcal{Y})$ a standard Borel space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes. Then there exists a schema $(\mu_n)_{n\in\mathbb{N}_0}$ for $\mathbf{y}$ given $\mathbf{x}$, where $\mu_n : H_n \times X \to \mathcal{P}(Y)$ is unique $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))$-a.e., for all $n \in \mathbb{N}_0$.*

Proof. According to Proposition A.5.16, since Y is a standard Borel space, each μ_n exists and is unique $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))$-a.e. □

By the uniqueness part of Proposition 3.1.1, it is possible to refer to *the* schema for $\mathbf{y}$ given $\mathbf{x}$.

A schema is noncontingent: for any history and any value in X, a schema determines the distribution on the possible values in Y.

Definition 3.1.2. A *schema base* is a finite set of schemas.

Let $H \triangleq \bigcup_{n\in\mathbb{N}_0} H_n$ be history space. The measurable function $\coprod_{n\in\mathbb{N}_0} \mu_n : H \times X \to \mathcal{P}(Y)$ is defined by $(\coprod_{n\in\mathbb{N}_0} \mu_n)|_{H_n\times X} = \mu_n$, for all $n \in \mathbb{N}_0$. This provides a convenient way of considering the sequence of functions $(\mu_n)_{n\in\mathbb{N}_0}$ all in one function.

Here is the second of the two main concepts of this book.

Definition 3.1.3. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ measurable spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes. An *empirical belief* (*for* $\mathbf{y}$ *given* $\mathbf{x}$) is a probability kernel

$$\lambda x.\mu_n(h, x) : X \to \mathcal{P}(Y),$$

where $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$ is a schema for $\mathbf{y}$ given $\mathbf{x}$ and $h \in H_n$, for some $n \in \mathbb{N}_0$.

By Proposition A.2.5, an empirical belief is indeed a probability kernel. However, there is no requirement that an empirical belief be a regular conditional distribution and, in fact, there cannot be such a requirement since there are many $\lambda x.\mu_n(h, x) : X \to \mathcal{P}(Y)$, each depending on the particular history h.

An empirical belief is contingent: given the specific history that has taken place, for any value in X, an empirical belief determines the distribution on the possible values in Y. The adjective 'empirical' is apt since an empirical belief depends upon the observations in the history. An empirical belief is a measure of an agent's state of knowledge about its environment, not necessarily a measure of the environment itself. Also, empirical beliefs are justified in the sense that they are instantiations using the current history of schemas which are regular conditional distributions.

Definition 3.1.4. An *empirical belief base* is a finite set of empirical beliefs.

As an example, an empirical belief base might consist of nothing more than the state distribution. Alternatively, an empirical belief base might contain a probability kernel that models (part of) the behaviour of some other agent. Intuitively, one can think of the state distribution as a model for the stage, and all the objects on it, on which actors perform. The actors are other agents and empirical beliefs may model aspects of their behaviour and/or beliefs.

Example 3.1.2. Consider a robot operating in some environment that includes a human with whom the robot must interact to achieve its goals. Let Z be the space of facial models of the person using features such as the shape of the mouth, whether the person is frowning ot not, how open the eyes are, and so on. The elements of Z can be obtained from observations. Let W be the space of mental states of the person. The robot can use the schema $(\mu_n : H_n \to \mathcal{P}(W^Z))_{n \in \mathbb{N}_0}$ to help decide which actions to perform to try to achieve its goal(s). Given the current history h_n, the empirical belief $\mu_n(h_n) : \mathcal{P}(W^Z)$ is a distribution over the function space W^Z. In this application, W is likely to be a finite set of classes, so that W^Z is an hypothesis space of classification functions. The distribution on $\mathcal{P}(W^Z)$ can then be used to help choose actions. For example, it is possible to extract a Bayes optimal classifier from W^Z using the distribution. Then, given a value for the facial model, the Bayes optimal classifier can compute a value for the mental state of the person. The discussion in Section 3.4 on function spaces, Example 3.4.5, and Example 4.3.1 show how distributions on function spaces can be used to do classification. Example 3.4.6 shows how distributions on function spaces can be used to do regresssion.

Now the structure of schemas, and their construction and deconstruction, is discussed. This is based on the structure of the space in the codomain that supports the probability measures. The two cases (for the forms of schema components) are:

- Products: $\mu_n : H_n \times X \to \mathcal{P}(\prod_{i \in I} Y_i)$
- Sums: $\mu_n : H_n \times X \to \mathcal{P}(\coprod_{i \in I} Y_i)$

In addition, for construction, there is a further case:

- Quotients: $\lambda x.(\mu_n(x) \circ p^{-1}) : H_n \times X \to \mathcal{P}(Z)$, where $\mu_n : H_n \times X \to \mathcal{P}(Y)$ and $p : Y \to Z$ is a measurable function

As explained in Section A.1, function spaces are regarded as special cases of product spaces. Thus components of the form $\mu_n : H_n \times X \to \mathcal{P}(W^Z)$ are included in the first case. It will be convenient to break the first case up into finite and infinite products. The next section considers the construction of schemas and the following section considers their deconstruction.

Note that, *before* deconstruction, 'top-level' schemas almost invariably have the form $(\mu_n : H_n \to \mathcal{P}(Y))_{n \in \mathcal{N}_0}$, where Y is a product space or a sum space. In particular, it is unlikely that 'top-level' schemas have the form $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n \in \mathcal{N}_0}$, with an argument X in the domain. The reason is that it generally makes much more sense for conditional situations to be modelled by a schema of the form $(\mu_n : H_n \to \mathcal{P}(Y^X))_{n \in \mathcal{N}_0}$ instead. However, *after* deconstruction, there will generally be many resulting schemas having the form $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n \in \mathcal{N}_0}$, as is shown in Section 3.3.

Finally, note that agents, environments, transition models, and observation models are not schemas; they strictly do not satisfy Definition 3.1.1. In particular, this means that if environments, transition models, or observation models need to be learned, this must be done by methods other than the stochastic filtering method of Chapter 4. In fact, transition models and observation models are needed to actually *perform* filtering.

3.2 Construction of Schemas

Here, the interest is in the construction of 'complex' schemas from 'simple' schemas. Each of the cases of products and sums is considered in turn.

3.2.1 Finite Products

This case is that of a finite product of schemas.

Consider schema components having signature

$$\mu_n^{(i)} : H_n \times X \times \prod_{j=1}^{i-1} Y_j \to \mathcal{P}(Y_i),$$

for $i = 1, \ldots, m$. The next proposition concerns the construction of the product schema with components having signature

$$\bigotimes_{i=1}^{m} \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\prod_{i=1}^{m} Y_i).$$

A detailed analysis of the product of probability kernels, including the propositions that the following proposition is essentially an immediate corollary of, is given in Section A.7.

Proposition 3.2.1. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y_i, \mathcal{Y}_i)$, for $i = 1, \ldots, m$, measurable spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y}^{(i)} : \Omega \to Y_i^{\mathbb{N}_0}$, for $i = 1, \ldots, m$, stochastic processes, and, for $i = 1, \ldots, m$, $(\mu_n^{(i)} : H_n \times X \times \prod_{j=1}^{i-1} Y_j \to \mathcal{P}(Y_i))_{n \in \mathbb{N}_0}$ a schema for $\mathbf{y}^{(i)}$ given $((\mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(i-1)}))_{n \in \mathbb{N}_0}$. Then the following hold.*

1. *$(\bigotimes_{i=1}^m \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\prod_{i=1}^m Y_i))_{n \in \mathbb{N}_0}$ is a schema for $\mathbf{y} \triangleq ((\mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(m)}))_{n \in \mathbb{N}_0}$ given $\mathbf{x}$.*
2. *For all $n \in \mathbb{N}_0$ and $h \in H_n$,*

$$\lambda x.(\bigotimes_{i=1}^{m} \mu_n^{(i)})(h, x) = \bigotimes_{i=1}^{m} \lambda(x, y_1, \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_1, \ldots, y_{i-1}).$$

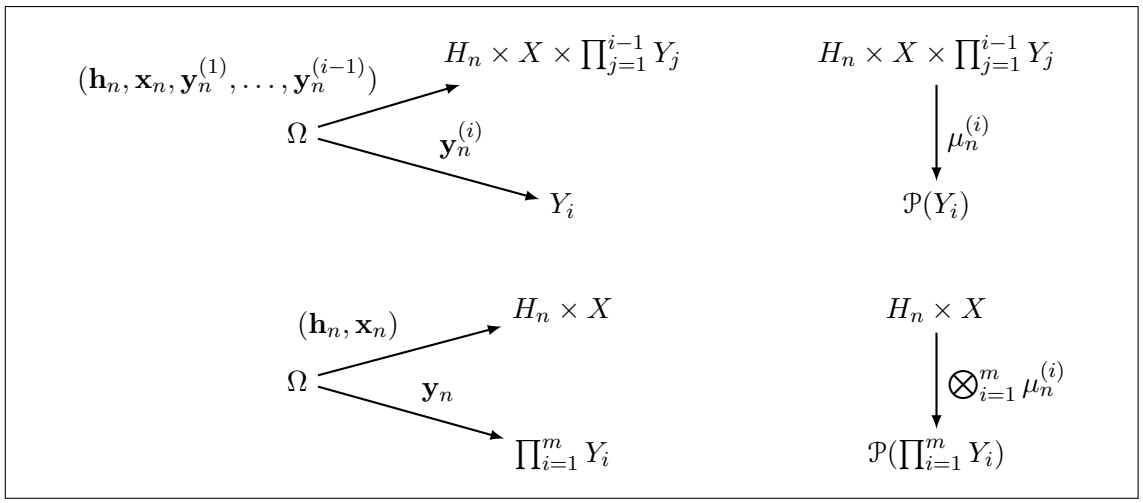

Figure 3.2: Setting for Proposition 3.2.1

Proof. 1. By Proposition A.1.5 and the measurability of each $\mathbf{y}^{(i)}$, $\mathbf{y} : \Omega \to (\prod_{i=1}^m Y_i)^{\mathbb{N}_0}$ is measurable. By Proposition A.7.3, $\bigotimes_{i=1}^m \mu_n^{(i)}$ is a probability kernel, for all $n \in \mathbb{N}_0$. Also, by Proposition A.7.12, $(\bigotimes_{i=1}^m \mu_n^{(i)})_{n \in \mathbb{N}_0}$ is a schema for $\mathbf{y}$ given $\mathbf{x}$.

2. Note that $\lambda x.(\bigotimes_{i=1}^m \mu_n^{(i)})(h, x)$ and each $\lambda(x, y_1, \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_1, \ldots, y_{i-1})$ is an empirical belief, by definition. This part follows directly from Proposition A.7.21. □

Proposition 3.2.1 is illustrated in Figures 3.3 and 3.4.

3.2.2 Infinite Products

This case is that of an infinite product of schemas.

Consider schema components having signature

$$\mu_n^{(m)} : H_n \times X \times \prod_{j=1}^{m-1} Y_j \to \mathcal{P}(Y_m),$$

$$\begin{cases} \mu_n^{(1)} : H_n \times X \to \mathcal{P}(Y_1) \\ \mu_n^{(2)} : H_n \times X \times Y_1 \to \mathcal{P}(Y_2) \\ \qquad\qquad \vdots \\ \mu_n^{(m)} : H_n \times X \times \prod_{i=1}^{m-1} Y_i \to \mathcal{P}(Y_m) \end{cases}$$

$$\Downarrow \quad [\textit{Proposition } 3.2.1]$$

$$\bigotimes_{i=1}^{m} \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\prod_{i=1}^{m} Y_i)$$

Figure 3.3: Finite product of schemas

$$\begin{cases} \lambda x.\mu_n^{(1)}(h,x) : X \to \mathcal{P}(Y_1) \\ \lambda(x,y_1).\mu_n^{(2)}(h,x,y_1) : X \times Y_1 \to \mathcal{P}(Y_2) \\ \qquad\qquad \vdots \\ \lambda(x,y_1,\ldots,y_{m-1}).\mu_n^{(m)}(h,x,y_1,\ldots,y_{m-1}) : X \times \prod_{i=1}^{m-1} Y_i \to \mathcal{P}(Y_m) \end{cases}$$

$$\Downarrow \quad [\textit{Proposition } 3.2.1]$$

$$\lambda x.(\bigotimes_{i=1}^{m} \mu_n^{(i)})(h,x) : X \to \mathcal{P}(\prod_{i=1}^{m} Y_i)$$

Figure 3.4: Finite product of empirical beliefs

for all $m \in \mathbb{N}$. The next proposition concerns the construction of the infinite product schema with components having signature

$$\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)} : H_n \times X \to \mathcal{P}(\prod_{m\in\mathbb{N}} Y_m).$$

A detailed analysis of the infinite product of probability kernels is given in Section A.8.

Proposition 3.2.2. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y_m, \mathcal{Y}_m)$, for all $m \in \mathbb{N}$, measurable spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y}^{(m)} : \Omega \to Y_m^{\mathbb{N}_0}$, for all $m \in \mathbb{N}$, stochastic processes, and, for all $m \in \mathbb{N}$, $(\mu_n^{(m)} : H_n \times X \times \prod_{j=1}^{m-1} Y_j \to \mathcal{P}(Y_m))_{n\in\mathbb{N}_0}$ a schema for $\mathbf{y}^{(m)}$ given $((\mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(m-1)}))_{n\in\mathbb{N}_0}$. Then the following hold.*

1. *$(\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)} : H_n \times X \to \mathcal{P}(\prod_{m\in\mathbb{N}} Y_m))_{n\in\mathbb{N}_0}$ is a schema for $\mathbf{y} \triangleq ((\mathbf{y}_n^{(1)}, \mathbf{y}_n^{(2)}, \ldots))_{n\in\mathbb{N}_0}$ given $\mathbf{x}$.*

2. For all $n \in \mathbb{N}_0$ and $h \in H_n$,

$$\lambda x.(\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)})(h,x) = \bigotimes_{m\in\mathbb{N}} \lambda(x, y_1, \ldots, y_{m-1}).\mu_n^{(m)}(h, x, y_1, \ldots, y_{m-1}).$$

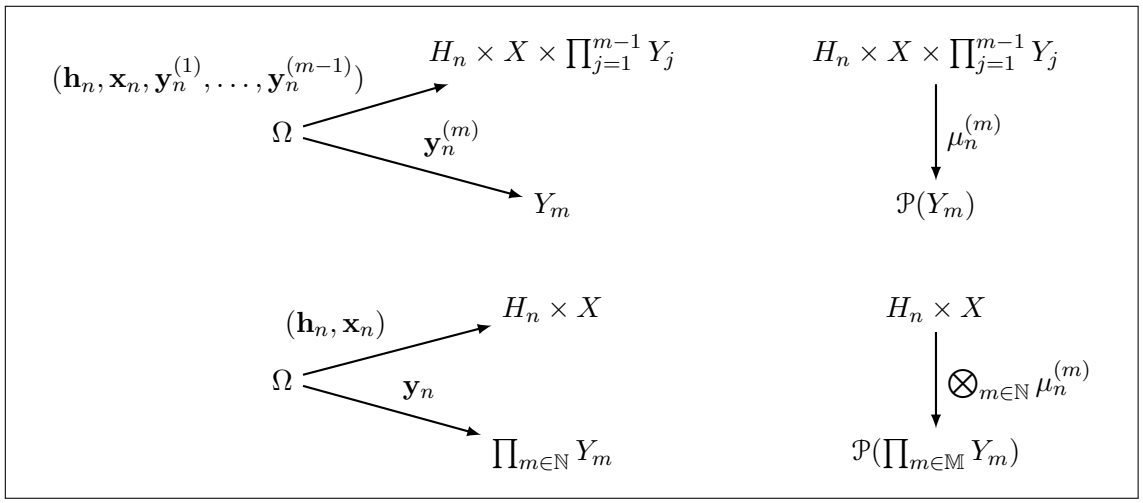

Figure 3.5: Setting for Proposition 3.2.2

Proof. 1. By Proposition A.1.5 and the measurability of each $\mathbf{y}^{(m)}$, $\mathbf{y} : \Omega \to (\prod_{m\in\mathbb{N}} Y_m)^{\mathbb{N}_0}$ is measurable. By Proposition A.8.3, $\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)}$ is a probability kernel, for all $n \in \mathbb{N}_0$. Also, by Proposition A.8.6, $(\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)})_{n\in\mathbb{N}_0}$ is a schema for $\mathbf{y}$ given $\mathbf{x}$.

2. Note that $\lambda x.(\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)})(h,x)$ and each $\lambda(x, y_1, \ldots, y_{m-1}).\mu_n^{(m)}(h, x, y_1, \ldots, y_{m-1})$ is an empirical belief, by definition. This part follows directly from Proposition A.8.5. □

Proposition 3.2.2 is illustrated in Figures 3.6 and 3.7.

3.2.3 Sums

This case is that of a sum of schemas.

Let I be a countable index set. Consider schema components having signature

$$\mu_n^{(i)} : H_n \times X \to \mathcal{P}(Y_i),$$

for $i \in I$. The goal is to produce a schema with components having signature

$$\bigoplus_{i\in I} \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\coprod_{i\in I} Y_i).$$

However, this turns out to be not possible in exactly this form. One possibility that could be shown to work using results in Section A.9 is instead to start with schema components having signature

$$\mu_n^{(i)} : H_n \times X \to \mathcal{P}(Y_i \sqcup \{*\}),$$

$$\begin{cases} \mu_n^{(1)} : H_n \times X \to \mathcal{P}(Y_1) \\ \mu_n^{(2)} : H_n \times X \times Y_1 \to \mathcal{P}(Y_2) \\ \qquad\qquad \vdots \\ \mu_n^{(m)} : H_n \times X \times \prod_{i=1}^{m-1} Y_i \to \mathcal{P}(Y_m) \\ \qquad\qquad \vdots \end{cases}$$

$$\Downarrow \quad [\textit{Proposition } 3.2.2]$$

$$\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)} : H_n \times X \to \mathcal{P}(\prod_{m\in\mathbb{N}} Y_m)$$

Figure 3.6: Infinite product of schemas

$$\begin{cases} \lambda x.\mu_n^{(1)}(h,x) : X \to \mathcal{P}(Y_1) \\ \lambda(x,y_1).\mu_n^{(2)}(h,x,y_1) : X \times Y_1 \to \mathcal{P}(Y_2) \\ \qquad\qquad \vdots \\ \lambda(x,y_1,\ldots,y_{m-1}).\mu_n^{(m)}(h,x,y_1,\ldots,y_{m-1}) : X \times \prod_{i=1}^{m-1} Y_i \to \mathcal{P}(Y_m) \\ \qquad\qquad \vdots \end{cases}$$

$$\Downarrow \quad [\textit{Proposition } 3.2.2]$$

$$\lambda x.(\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)})(h,x) : X \to \mathcal{P}(\prod_{m\in\mathbb{N}} Y_m)$$

Figure 3.7: Infinite product of empirical beliefs

for $i \in I$. Here, $Y_i \sqcup \{*\}$ is the one point extension of Y discussed in Sections A.1 and A.9. Then the $\bigoplus_{i\in I} \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\coprod_{i\in I} Y_i)$ can be successfully defined. However, this approach is not pursued here since the $\mu_n^{(i)} : H_n \times X \to \mathcal{P}(Y_i \sqcup \{*\})$ generally cannot be deconstructed – a significant defect – due to the presence of the $\{*\}$.

Instead, from the $\mu_n^{(i)} : H_n \times X \to \mathcal{P}(Y_i)$, a probability kernel that is a weighted sum and whose components have signature

$$\breve{\chi}_n \bullet \bigoplus_{i\in I} \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\coprod_{i\in I} Y_i)$$

is defined. Here is the relevant proposition that follows directly from results in Section A.9.

Proposition 3.2.3. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(I, \mathbb{B}^I, c)$ a countable measure space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y_i, \mathcal{Y}_i)$, for all $i \in I$, measurable spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{i} : \Omega \to I^{\mathbb{N}_0}$ stochastic processes, $\mathbf{y}^{(i)} : \Omega \to (Y_i)^{\mathbb{N}_0}$ stochastic processes, for all $i \in I$, $(\chi_n : H_n \times X \to \mathcal{P}(I))_{n\in\mathbb{N}_0}$ a schema for $\mathbf{i}$ given $\mathbf{x}$, and $(\mu_n^{(i)} : H_n \times X \to \mathcal{P}(Y_i))_{n\in\mathbb{N}_0}$ a schema for $\mathbf{y}^{(i)}$ given $\mathbf{x}$, for all $i \in I$. Define $\breve{\chi}_n : H_n \times X \to \mathcal{D}(I)$ by $\breve{\chi}_n(h, x)(i) = \chi_n(h, x)(\{i\})$, for all $x \in X$, $h \in H_n$, and $i \in I$. Define the stochastic process $\mathbf{y} : \Omega \to (\coprod_{i\in I} Y_i)^{\mathbb{N}_0}$ by $\mathbf{y} = (\mathbf{y}_n)_{n\in\mathbb{N}_0}$, where $\mathbf{y}_n : \Omega \to \coprod_{i\in I} Y_i$ is defined by $\mathbf{y}_n = \lambda\omega.\mathbf{y}_n^{(\mathbf{i}_n(\omega))}(\omega)$, for all $n \in \mathbb{N}_0$. Suppose that $\sigma(\mathbf{i}_n)$ and $\sigma(\mathbf{y}_n^{(i)})$ are conditionally independent given $\sigma(\mathbf{h}_n, \mathbf{x}_n)$, for all $i \in I$ and $n \in \mathbf{N}_0$. Then the following hold.*

1. *$(\breve{\chi}_n \bullet \bigoplus_{i\in I} \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\coprod_{i\in I} Y_i))_{n\in\mathbb{N}_0}$, is a schema for $\mathbf{y}$ given $\mathbf{x}$.*
2. *For all $n \in \mathbb{N}_0$ and $h \in H_n$,*

$$\lambda x.(\breve{\chi}_n \bullet \bigoplus_{i\in I} \mu_n^{(i)})(h, x) = \lambda x.\breve{\chi}_n(h, x) \bullet \bigoplus_{i\in I} \lambda x.\mu_n^{(i)}(h, x).$$

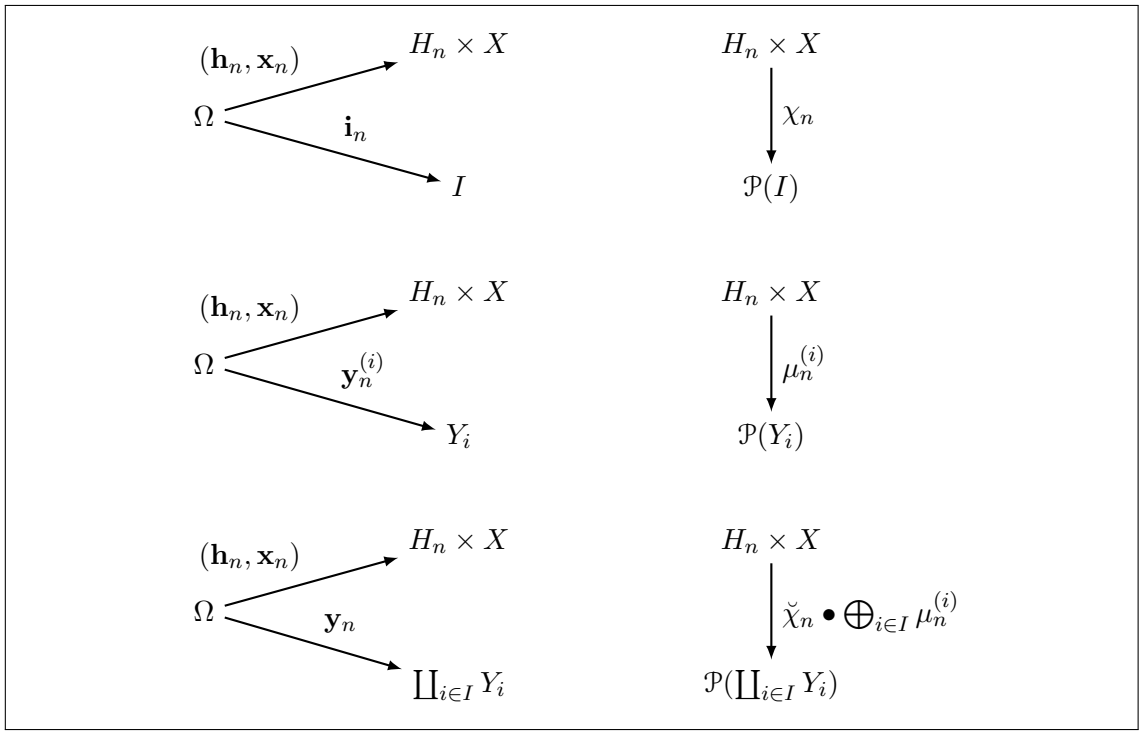

Figure 3.8: Setting for Proposition 3.2.3

Proof. 1. This part follows directly from Proposition A.9.10.

2. Note that $\lambda x.(\breve{\chi}_n \bullet \bigoplus_{i\in I} \mu_n^{(i)})(h, x)$ and each $\lambda x.\mu_n^{(i)}(h, x)$ is an empirical belief, by definition. This part follows directly from Proposition A.9.12. □

Proposition 3.2.3 is illustrated in Figures 3.9 and 3.10.

$$\begin{cases} \chi_n : H_n \times X \to \mathcal{P}(I) \\ \mu_n^{(i_1)} : H_n \times X \to \mathcal{P}(Y_{i_1}) \\ \mu_n^{(i_2)} : H_n \times X \to \mathcal{P}(Y_{i_2}) \\ \qquad\qquad \vdots \\ \mu_n^{(i_m)} : H_n \times X \to \mathcal{P}(Y_{i_m}) \\ \qquad\qquad \vdots \end{cases}$$

$$\Downarrow \quad [\textit{Proposition } 3.2.3]$$

$$\breve{\chi}_n \bullet \bigoplus_{i \in I} \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\coprod_{i \in I} Y_i)$$

Figure 3.9: Sum of schemas

$$\begin{cases} \lambda x.\chi_n(h,x) : X \to \mathcal{P}(I) \\ \lambda x.\mu_n^{(i_1)}(h,x) : X \to \mathcal{P}(Y_{i_1}) \\ \lambda x.\mu_n^{(i_2)}(h,x) : X \to \mathcal{P}(Y_{i_2}) \\ \qquad\qquad \vdots \\ \lambda x.\mu_n^{(i_m)}(h,x) : X \to \mathcal{P}(Y_{i_m}) \\ \qquad\qquad \vdots \end{cases}$$

$$\Downarrow \quad [\textit{Proposition } 3.2.3]$$

$$\lambda x.\breve{\chi}_n(h,x) \bullet \bigoplus_{i \in I} \lambda x.\mu_n^{(i)}(h,x) : X \to \mathcal{P}(\coprod_{i \in I} Y_i)$$

Figure 3.10: Sum of empirical beliefs

3.3 Deconstruction of Schemas

When modelling for an application, there is an interest in deconstructing 'complex' schemas into 'simple' schemas, which may be easier to acquire and utilize. Thus this section considers the deconstruction of schemas. The relevant cases are products, sums, and quotients.

3.3.1 Finite Products

First it is shown how schemas in the finite product case can be deconstructed.

Proposition 3.3.1. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ a measurable space, $(Y_i, \mathcal{Y}_i)$, for $i = 1, \ldots, m$, a standard Borel space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to (\prod_{i=1}^m Y_i)^{\mathbb{N}_0}$ stochastic processes, and $(\mu_n : H_n \times X \to \mathcal{P}(\prod_{i=1}^m Y_i))_{n\in\mathbb{N}_0}$ a schema for $\mathbf{y}$ given $\mathbf{x}$. Then the following hold.*

1. *For $i = 1, \ldots, m$, there exists a stochastic process $\mathbf{y}^{(i)} : \Omega \to Y_i^{\mathbb{N}_0}$ and a schema $(\mu_n^{(i)} : H_n \times X \times \prod_{j=1}^{i-1} Y_j \to \mathcal{P}(Y_i))_{n\in\mathbb{N}_0}$ for $\mathbf{y}^{(i)}$ given $((\mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(i-1)}))_{n\in\mathbb{N}_0}$ such that*

$$\mu_n = \bigotimes_{i=1}^m \mu_n^{(i)} \quad \mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))\text{-a.e.},$$

   for all $n \in \mathbb{N}_0$.

2. *For all $n \in \mathbb{N}_0$ and $h \in H_n$,*

$$\lambda x.\mu_n(h, x) = \bigotimes_{i=1}^m \lambda(x, y_1, \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_1, \ldots, y_{i-1}),$$

   except on $\{x \in X \mid (h, x) \in N_n\}$, where $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))(N_n) = 0$.

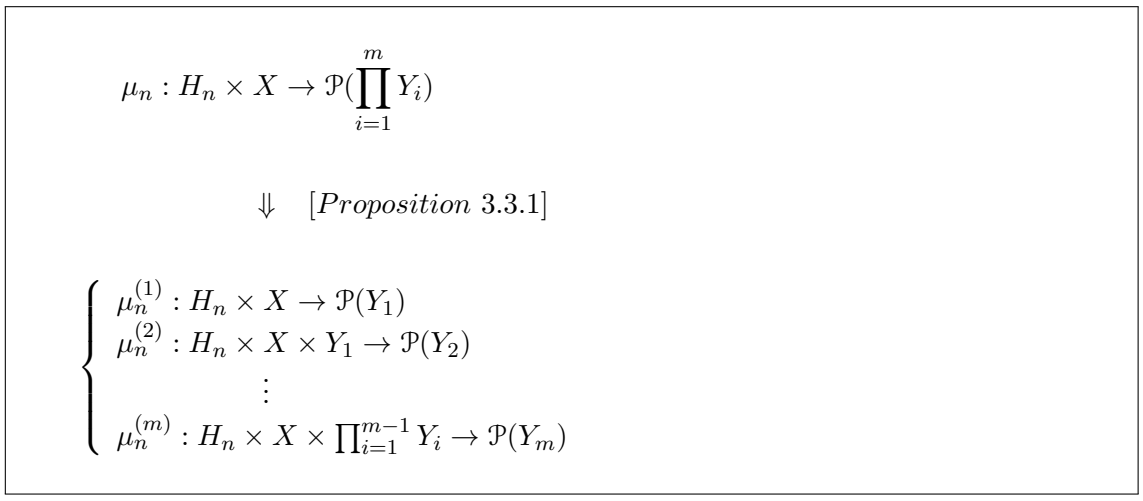

Figure 3.11: Deconstruction of schemas in the finite product case

$$\lambda x.\mu_n(h,x) : X \to \mathcal{P}(\prod_{i=1}^{m} Y_i)$$

$$\Downarrow \quad [\textit{Proposition } 3.3.1]$$

$$\begin{cases} \lambda x.\mu_n^{(1)}(h,x) : X \to \mathcal{P}(Y_1) \\ \lambda(x,y_1).\mu_n^{(2)}(h,x,y_1) : X \times Y_1 \to \mathcal{P}(Y_2) \\ \qquad\qquad \vdots \\ \lambda(x,y_1,\ldots,y_{m-1}).\mu_n^{(m)}(h,x,y_1,\ldots,y_{m-1}) : X \times \prod_{i=1}^{m-1} Y_i \to \mathcal{P}(Y_m) \end{cases}$$

Figure 3.12: Deconstruction of empirical beliefs in the finite product case

Proof. 1. For $i = 1,\ldots,m$, define $\mathbf{y}^{(i)} : \Omega \to Y_i^{\mathbb{N}_0}$ by $\mathbf{y}^{(i)} = (\mathbf{y}_n^{(i)})_{n\in\mathbb{N}_0}$, where $\mathbf{y}_n^{(i)} : \Omega \to Y_i$ is defined by $\mathbf{y}_n^{(i)} = \lambda(y_1,\ldots,y_m).y_i \circ \mathbf{y}_n$, for all $n \in \mathbb{N}_0$. Clearly, each $\mathbf{y}^{(i)} : \Omega \to Y_i^{\mathbb{N}_0}$ is a stochastic process. For all $n \in \mathbb{N}_0$, by Proposition A.7.19, there exists a probability kernel $\mu_n^{(i)} : H_n \times X \times \prod_{j=1}^{i-1} Y_j \to \mathcal{P}(Y_i)$ that is a regular conditional distribution of $\mathbf{y}_n^{(i)}$ given $(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(1)},\ldots,\mathbf{y}_n^{(i-1)})$, for $i = 1,\ldots,m$, such that $\mu_n = \bigotimes_{i=1}^m \mu_n^{(i)}$ $\mathcal{L}((\mathbf{h}_n,\mathbf{x}_n))$-a.e. Also $(\mu_n^{(i)})_{n\in\mathbb{N}_0}$ is a schema for $\mathbf{y}^{(i)}$ given $((\mathbf{x}_n,\mathbf{y}_n^{(1)},\ldots,\mathbf{y}_n^{(i-1)}))_{n\in\mathbb{N}_0}$, for $i = 1,\ldots,m$.

2. Let $N_n \subseteq H_n \times X$ be such that $\mathcal{L}((\mathbf{h}_n,\mathbf{x}_n))(N_n) = 0$ and $\mu_n = \bigotimes_{i=1}^m \mu_n^{(i)}$ on $(H_n \times X) \setminus N_n$. By Proposition A.7.21, for all $n \in \mathbb{N}_0$ and $h \in H_n$,

$$\lambda x.\mu_n(h,x) = \bigotimes_{i=1}^{m} \lambda(x,y_1\ldots,y_{i-1}).\mu_n^{(i)}(h,x,y_1,\ldots,y_{i-1}),$$

except on $\{x \in X \mid (h,x) \in N_n\}$. □

Next is an important generalization of Proposition 3.3.1 that takes conditional independence conditions into account.

Notation. For $i = 1,\ldots,n$, suppose that $par(i) \triangleq \{i_1,\ldots,i_m\} \subseteq \{1,\ldots,i-1\}$, where $i_1 < \cdots < i_m$. Then $\mathbf{y}_n^{(par(i))}$ denotes $(\mathbf{y}_n^{(i_1)},\ldots,\mathbf{y}_n^{(i_m)})$. Similarly, $y_{par(i)}$ denotes $(y_{i_1},\ldots,y_{i_m})$.

Proposition 3.3.2. *Let $(\Omega,\mathfrak{S},\mathsf{P})$ be a probability space, $(A,\mathcal{A})$ an action space, $(O,\mathcal{O})$ an observation space, $(X,\mathcal{X})$ a measurable space, $(Y_i,\mathcal{Y}_i)$, for $i = 1,\ldots,m$, a standard Borel space $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to (\prod_{i=1}^m Y_i)^{\mathbb{N}_0}$ stochastic processes, and $(\mu_n : H_n \times X \to \mathcal{P}(\prod_{i=1}^m Y_i))_{n\in\mathbb{N}_0}$ a schema for $\mathbf{y}$ given $\mathbf{x}$. For $i = 1,\ldots,m$, define $\mathbf{y}^{(i)} : \Omega \to Y_i^{\mathbb{N}_0}$ by $\mathbf{y}^{(i)} = (\mathbf{y}_n^{(i)})_{n\in\mathbb{N}_0}$, where $\mathbf{y}_n^{(i)} : \Omega \to Y_i$ is defined by $\mathbf{y}_n^{(i)} = \lambda(y_1,\ldots,y_m).y_i \circ \mathbf{y}_n$, for all $n \in \mathbb{N}_0$. Suppose that there is a dependency graph with vertices $1,\ldots,m$, where each vertex i is labelled by $\sigma(\mathbf{y}_n^{(i)})$. Suppose also that, for $i = 1,\ldots,m$,*

$$\sigma(\mathbf{y}_n^{(i)}) \underset{\sigma(\mathbf{h}_n,\mathbf{x}_n,\mathbf{y}_n^{(par(i))})}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n,\mathbf{x}_n,\mathbf{y}_n^{(1)},\ldots,\mathbf{y}_n^{(i-1)}),$$

for all $n \in \mathbb{N}_0$. Then the following hold.

1. *For $i = 1, \ldots, m$, there exists a schema $(\mu_n^{(i)} : H_n \times X \times \prod_{j \in par(i)} Y_j \to \mathcal{P}(Y_i))_{n \in \mathbb{N}_0}$ for $\mathbf{y}^{(i)}$ given $((\mathbf{x}_n, \mathbf{y}_n^{(par(i))}))_{n \in \mathbb{N}_0}$ such that*

$$\mu_n = \bigotimes_{i=1}^{m} \lambda(h, x, y_1, \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_{par(i)}) \;\; \mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))\text{-a.e.},$$

   for all $n \in \mathbb{N}_0$. Furthermore, $(\lambda(h, x, y_1, \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_{par(i)}))_{n \in \mathbb{N}_0}$ is a schema for $\mathbf{y}^{(i)}$ given $((\mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(i-1)}))_{n \in \mathbb{N}_0}$, for $i = 1, \ldots, m$.

2. *For all $n \in \mathbb{N}_0$ and $h \in H_n$,*

$$\lambda x.\mu_n(h, x) = \bigotimes_{i=1}^{m} \lambda(x, y_1, \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_{par(i)}),$$

   except on $\{x \in X \mid (h, x) \in N_n\}$, where $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))(N_n) = 0$.

Proof. 1. Clearly each $\mathbf{y}^{(i)} : \Omega \to Y_i^{\mathbb{N}_0}$ is a stochastic process. For all $n \in \mathbb{N}_0$, by Proposition A.7.22, there exists a probability kernel $\mu_n^{(i)} : H_n \times X \times \prod_{j \in par(i)} Y_j \to \mathcal{P}(Y_i)$ that is a regular conditional distribution of $\mathbf{y}_n^{(i)}$ given $(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(par(i))})$, for $i = 1, \ldots, m$, such that $\mu_n = \bigotimes_{i=1}^{m} \lambda(h, x, y_1, \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_{par(i)}) \;\; \mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))$-a.e. Proposition A.7.22 also shows that, for $i = 1, \ldots, m$,

$$(\lambda(h, x, y_1, \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_{par(i)}) : H_n \times X \times \prod_{j=1}^{i-1} Y_j \to \mathcal{P}(Y_i))_{n \in \mathbb{N}_0}$$

is a schema for $\mathbf{y}^{(i)}$ given $((\mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(i-1)}))_{n \in \mathbb{N}_0}$.

2. Let $N_n \subseteq H_n \times X$ be such that $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))(N_n) = 0$ and $\mu_n = \bigotimes_{i=1}^{m} \mu_n^{(i)}$ on $(H_n \times X) \setminus N_n$. By Proposition A.7.21, for all $n \in \mathbb{N}_0$ and $h \in H_n$,

$$\lambda x.\mu_n(h, x) = \bigotimes_{i=1}^{m} \lambda(x, y_1 \ldots, y_{i-1}).\mu_n^{(i)}(h, x, y_{par(i)}),$$

except on $\{x \in X \mid (h, x) \in N_n\}$. □

By Proposition A.6.1, the condition

$$\sigma(\mathbf{y}_n^{(i)}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(par(i))})}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(i-1)})$$

is equivalent to

$$\mathsf{P}((\mathbf{y}_n^{(i)})^{-1}(B_i) \mid (\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(i-1)})) = \mathsf{P}((\mathbf{y}_n^{(i)})^{-1}(B_i) \mid (\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(par(i))})),$$

for all $B_i \in Y_i$. In this form, it is perhaps clearer that the conditional independence assumptions in Proposition 3.3.2 do accurately capture our intuition about the intended meaning of these assumptions. Note that these conditional independence assumptions are *persistent* in the sense that they hold for all $n \in \mathbb{N}_0$, that is, they hold throughout the lifetime of the agent employing the corresponding schema.

3.3.2 Infinite Products

The next case is when the codomain of the schema is the set of probability measures on an infinite product space. This includes the special cases of probability measures on sets and also multisets.

Proposition 3.3.3. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ a measurable space, $(Y_m, \mathcal{Y}_m)$, for all $m \in \mathbb{N}$, a standard Borel space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to (\prod_{m\in\mathbb{N}} Y_m)^{\mathbb{N}_0}$ stochastic processes, and $(\mu_n : H_n \times X \to \mathcal{P}(\prod_{m\in\mathbb{N}} Y_m))_{n\in\mathbb{N}_0}$ a schema for $\mathbf{y}$ given $\mathbf{x}$. Then the following hold.*

1. *For all $m \in \mathbb{N}$, there exists a stochastic process $\mathbf{y}^{(m)} : \Omega \to Y_m^{\mathbb{N}_0}$ and a schema $(\mu_n^{(m)} : H_n \times X \times \prod_{j=1}^{m-1} Y_j \to \mathcal{P}(Y_m))_{n\in\mathbb{N}_0}$ for $\mathbf{y}^{(m)}$ given $((\mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(m-1)}))_{n\in\mathbb{N}_0}$ such that*

$$\mu_n = \bigotimes_{m\in\mathbb{N}} \mu_n^{(m)} \quad \mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))\text{-}a.e.,$$

   for all $n \in \mathbb{N}_0$.

2. *For all $n \in \mathbb{N}_0$ and $h \in H_n$,*

$$\lambda x.\mu_n(h, x) = \bigotimes_{m\in\mathbb{N}} \lambda(x, y_1, \ldots, y_{m-1}).\mu_n^{(m)}(h, x, y_1, \ldots, y_{m-1}),$$

   except on $\{x \in X \mid (h, x) \in N_n\}$, where $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))(N_n) = 0$.

$$\mu_n : H_n \times X \to \mathcal{P}(\prod_{m\in\mathbb{N}} Y_m)$$

$$\Downarrow \quad [\textit{Proposition } 3.3.3]$$

$$\begin{cases} \mu_n^{(1)} : H_n \times X \to \mathcal{P}(Y_1) \\ \mu_n^{(2)} : H_n \times X \times Y_1 \to \mathcal{P}(Y_2) \\ \qquad\vdots \\ \mu_n^{(m)} : H_n \times X \times \prod_{i=1}^{m-1} Y_i \to \mathcal{P}(Y_m) \\ \qquad\vdots \end{cases}$$

Figure 3.13: Deconstruction of schemas in the infinite product case

Proof. 1. For all $m \in \mathbb{N}$, define $\mathbf{y}^{(m)} : \Omega \to Y_m^{\mathbb{N}_0}$ by $\mathbf{y}^{(m)} = (\mathbf{y}_n^{(m)})_{n\in\mathbb{N}_0}$, where $\mathbf{y}_n^{(m)} : \Omega \to Y_m$ is defined by $\mathbf{y}_n^{(m)} = \lambda(y_1, y_2, \ldots).y_m \circ \mathbf{y}_n$, for all $n \in \mathbb{N}_0$. Clearly, each $\mathbf{y}^{(m)} : \Omega \to Y_m^{\mathbb{N}_0}$

$$\lambda x.\mu_n(h,x) : X \to \mathcal{P}(\prod_{m\in\mathbb{N}} Y_m)$$

$$\Downarrow \quad [\textit{Proposition } 3.3.3]$$

$$\begin{cases} \lambda x.\mu_n^{(1)}(h,x) : X \to \mathcal{P}(Y_1) \\ \lambda(x,y_1).\mu_n^{(2)}(h,x,y_1) : X \times Y_1 \to \mathcal{P}(Y_2) \\ \qquad\vdots \\ \lambda(x,y_1,\ldots,y_{m-1}).\mu_n^{(m)}(h,x,y_1,\ldots,y_{m-1}) : X \times \prod_{i=1}^{m-1} Y_i \to \mathcal{P}(Y_m) \\ \qquad\vdots \end{cases}$$

Figure 3.14: Deconstruction of empirical beliefs in the infinite product case

is a stochastic process. For all $n \in \mathbb{N}_0$, by Proposition A.8.8, there exists a probability kernel $\mu_n^{(m)} : H_n \times X \times \prod_{j=1}^{m-1} Y_j \to \mathcal{P}(Y_m)$ that is a regular conditional distribution of $\mathbf{y}_n^{(m)}$ given $(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(m-1)})$, for all $m \in \mathbb{N}$, such that $\mu_n = \bigotimes_{m\in\mathbb{N}} \mu_n^{(m)}$ $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))$-a.e. Also $(\mu_n^{(m)})_{n\in\mathbb{N}_0}$ is a schema for $\mathbf{y}^{(m)}$ given $((\mathbf{x}_n, \mathbf{y}_n^{(1)}, \ldots, \mathbf{y}_n^{(m-1)}))_{n\in\mathbb{N}_0}$, for all $m \in \mathbb{N}$.

2. Let $N_n \subseteq H_n \times X$ be such that $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))(N_n) = 0$ and $\mu_n = \bigotimes_{m\in\mathbb{N}} \mu_n^{(m)}$ on $(H_n \times X) \setminus N_n$. By Proposition A.8.5, for all $h \in H_n$,

$$\lambda x.\mu_n(h,x) = \bigotimes_{m\in\mathbb{N}} \lambda(x, y_1 \ldots, y_{m-1}).\mu_n^{(m)}(h,x,y_1,\ldots,y_{m-1}),$$

except on $\{x \in X \mid (h,x) \in N_n\}$. □

3.3.3 Sums

This case is when the codomain of the schema is the set of probability measures on a sum space.

Two deconstructions are presented. The first deconstructs an *arbitrary* schema $(\mu_n : H_n \times X \to \mathcal{P}(\coprod_{i\in I} Y_i))_{n\in\mathbb{N}_0}$.

Proposition 3.3.4. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y_i, \mathcal{Y}_i)$, for all $i \in I$, measurable spaces, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to (\coprod_{i\in I} Y_i)^{\mathbb{N}_0}$ stochastic processes, and $(\mu_n : H_n \times X \to \mathcal{P}(\coprod_{i\in I} Y_i))_{n\in\mathbb{N}_0}$ a schema for $\mathbf{y}$ given $\mathbf{x}$. Then the following hold.*

1. *For all $i \in I$, there exists a stochastic process $\mathbf{y}^{(i)} : \Omega \to (Y_i \sqcup \{*\})^{\mathbb{N}_0}$ and a schema $(\mu_n^{(i)} : H_n \times X \to \mathcal{P}(Y_i \sqcup \{*\}))_{n\in\mathbb{N}_0}$ for $\mathbf{y}^{(i)}$ given $\mathbf{x}$ such that*

$$\mu_n = \bigoplus_{i\in I} \mu_n^{(i)},$$

for all $n \in \mathbb{N}_0$.

Furthermore, $\sum_{i \in I} \mu_n^{(i)}(h,x)(Y_i) = 1$, *for all* $h \in H_n$, $x \in X$, *and* $n \in \mathbb{N}_0$.

2. *For all* $n \in \mathbb{N}_0$ *and* $h \in H_n$,

$$\lambda x.\mu_n(h,x) = \bigoplus_{i \in I} \lambda x.\mu_n^{(i)}(h,x).$$

$$\mu_n : H_n \times X \to \mathcal{P}(\coprod_{i \in I} Y_i)$$

$$\Downarrow \quad [\textit{Proposition } 3.3.4]$$

$$\begin{cases} \mu_n^{(i_1)} : H_n \times X \to \mathcal{P}(Y_{i_1} \sqcup \{*\}) \\ \mu_n^{(i_2)} : H_n \times X \to \mathcal{P}(Y_{i_2} \sqcup \{*\}) \\ \qquad\qquad \vdots \\ \mu_n^{(i_m)} : H_n \times X \to \mathcal{P}(Y_{i_m} \sqcup \{*\}) \\ \qquad\qquad \vdots \end{cases}$$

Figure 3.15: Deconstruction of schemas in the sum case

$$\lambda x.\mu_n(h,x) : X \to \mathcal{P}(\coprod_{i \in I} Y_i)$$

$$\Downarrow \quad [\textit{Proposition } 3.3.4]$$

$$\begin{cases} \lambda x.\mu_n^{(i_1)}(h,x) : X \to \mathcal{P}(Y_{i_1} \sqcup \{*\}) \\ \lambda x.\mu_n^{(i_2)}(h,x) : X \to \mathcal{P}(Y_{i_2} \sqcup \{*\}) \\ \qquad\qquad \vdots \\ \lambda x.\mu_n^{(i_m)}(h,x) : X \to \mathcal{P}(Y_{i_m} \sqcup \{*\}) \\ \qquad\qquad \vdots \end{cases}$$

Figure 3.16: Deconstruction of empirical beliefs in the sum case

Proof. 1. For all $i \in I$, define $\mathbf{y}^{(i)} : \Omega \to (Y_i \sqcup \{*\})^{\mathbb{N}_0}$ by $\mathbf{y}^{(i)} = (\mathbf{y}_n^{(i)})_{n \in \mathbb{N}_0}$, where

$\mathbf{y}_n^{(i)} : \Omega \to Y_i \sqcup \{*\}$ is defined by

$$\mathbf{y}_n^{(i)}(\omega) = \begin{cases} \mathbf{y}_n(\omega) & \text{if } \omega \in \mathbf{y}_n^{-1}(Y_i) \\ * & \text{otherwise,} \end{cases}$$

for all $\omega \in \Omega$. Clearly, each $\mathbf{y}_i$ is a stochastic process. By Proposition A.9.4, for all $i \in I$, there exists a schema $(\mu_n^{(i)} : H_n \times X \to \mathcal{P}(Y_i \sqcup \{*\}))_{n\in\mathbb{N}_0}$ for $\mathbf{y}^{(i)}$ given $\mathbf{x}$ such that $\mu_n = \bigoplus_{i\in I} \mu_n^{(i)}$. Proposition A.9.2 shows that $\sum_{i\in I} \mu_n^{(i)}(h,x)(Y_i) = 1$, for all $h \in H_n$, $x \in X$, and $n \in \mathbb{N}_0$.

2. For all $n \in \mathbb{N}_0$, $\sum_{i\in I} \mu_n^{(i)}(h,x)(Y_i) = 1$, for all $h \in H_n$ and $x \in X$. Hence, by Proposition A.9.5, for all $n \in \mathbb{N}_0$ and $h \in H_n$, $\lambda x.\mu_n(h,x) = \bigoplus_{i\in I} \lambda x.\mu_n^{(i)}(h,x)$. □

A problem with the above deconstruction is the introduction of the set $\{*\}$ that means that, generally, further deconstruction of each $\mu_n^{(i)}$ is blocked. An alternative deconstruction depends on restricting the class of schemas to be considered. More precisely, consider those schema that have the form of a weighted sum $\breve{\chi}_n \bullet \bigoplus_{i\in I} \mu_n^{(i)}$, for some χ and $\mu^{(i)}$, for all $i \in I$ (as in Proposition 3.2.3). This restriction still provides a rich class of schemas suitable for applications and means that further deconstruction can take place by deconstructing the $\mu_n^{(i)}$. It also means that the deconstruction of each schema in the class is trivial, by definition. This is indicated in Figures 3.17 and 3.18.

$$\breve{\chi}_n \bullet \bigoplus_{i\in I} \mu_n^{(i)} : H_n \times X \to \mathcal{P}(\coprod_{i\in I} Y_i)$$

$$\Downarrow$$

$$\begin{cases} \chi_n : H_n \times X \to \mathcal{P}(I) \\ \mu_n^{(i_1)} : H_n \times X \to \mathcal{P}(Y_{i_1}) \\ \mu_n^{(i_2)} : H_n \times X \to \mathcal{P}(Y_{i_2}) \\ \qquad\vdots \\ \mu_n^{(i_m)} : H_n \times X \to \mathcal{P}(Y_{i_m}) \\ \qquad\vdots \end{cases}$$

Figure 3.17: Deconstruction of schemas in the weighted sum case

3.3.4 Quotients

This is the case of a quotient schema.

Consider schema components having signature

$$\mu_n : H_n \times X \to \mathcal{P}(Y)$$

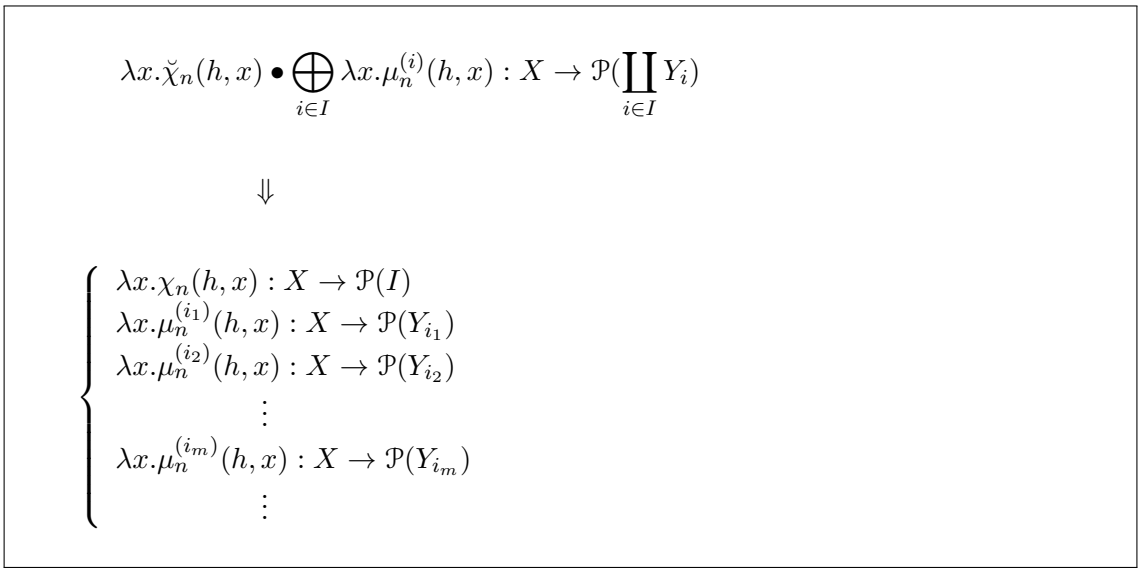

Figure 3.18: Deconstruction of empirical beliefs in the weighted sum case

and let $p : Y \to Z$ be a measurable function. The next proposition concerns the construction of the quotient schema with components having signature

$$\mu_n/p : H_n \times X \to \mathcal{P}(Z).$$

A detailed analysis of the quotient of a probability kernel is given in Section A.10.

Proposition 3.3.5. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$, $(Y, \mathcal{Y})$, and $(Z, \mathcal{Z})$ measurable spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes, $p : Y \to Z$ a measurable function, and $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ a schema for $\mathbf{y}$ given $\mathbf{x}$. Then the following hold.*

1. *$(\mu_n/p : H_n \times X \to \mathcal{P}(Z))_{n\in\mathbb{N}_0}$ is a schema for $(p \circ \mathbf{y}_n)_{n\in\mathbb{N}_0}$ given $\mathbf{x}$.*

2. *For all $n \in \mathbb{N}_0$ and $h \in H_n$,*

$$\lambda x.(\mu_n/p)(h,x) = \lambda x.\mu_n(h,x)/p.$$

Proof. 1. By Proposition A.10.2, each μ_n/p is a regular conditional distribution of $p \circ \mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{x}_n)$. Hence $(\mu_n/p)_{n\in\mathbb{N}_0}$ is a schema for $(p \circ \mathbf{y}_n)_{n\in\mathbb{N}_0}$ given $\mathbf{x}$.

2. Note that $\lambda x.(\mu_n/p)(h,x)$ and $\lambda x.\mu_n(h,x)$ are empirical beliefs, by definition. This part follows directly from Proposition A.10.3. □

Proposition 3.3.5 is illustrated in Figures 3.20 and 3.21.

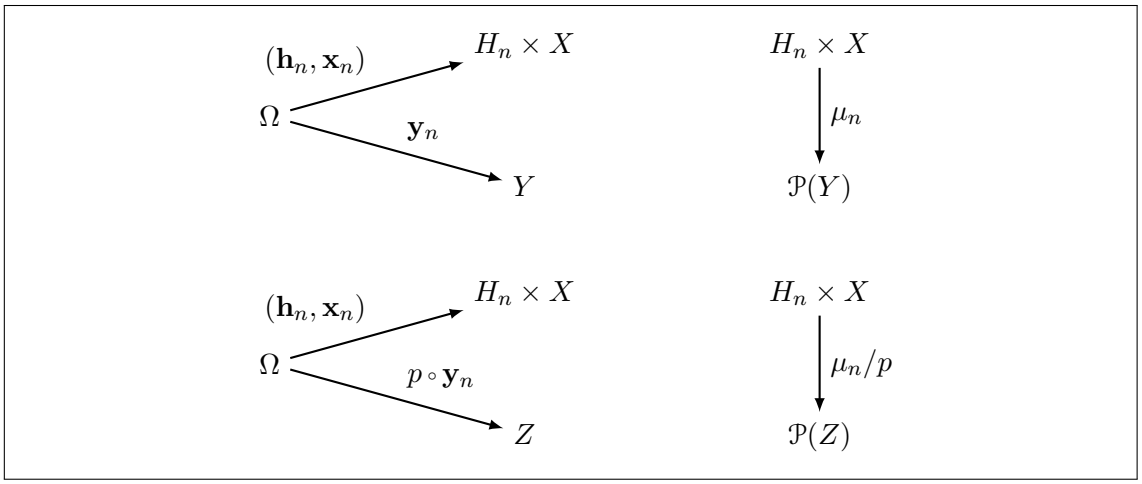

Figure 3.19: Setting for Proposition 3.3.5

$$\begin{cases} \mu_n : H_n \times X \to \mathcal{P}(Y) \\ p : Y \to Z \end{cases}$$

$$\Downarrow \quad [\textit{Proposition } 3.3.5]$$

$$\mu_n / p : H_n \times X \to \mathcal{P}(Z)$$

Figure 3.20: Quotient of a schema

$$\begin{cases} \lambda x.\mu_n(h, x) : X \to \mathcal{P}(Y) \\ p : Y \to Z \end{cases}$$

$$\Downarrow \quad [\textit{Proposition } 3.3.5]$$

$$\lambda x.(\mu_n / p)(h, x) : X \to \mathcal{P}(Z)$$

Figure 3.21: Quotient of an empirical belief

3.3.5 Deconstruction Examples

This subsection contains two examples to illustrate how a schema can be deconstructed into schemas whose codomains are the set of all probability measures over 'simple' spaces, such as finite sets, $\mathbb{N}_0$, or $\mathbb{R}^k$, for some k.

Example 3.3.1. Consider a schema whose codomain is the set of probability measures on lists of 5-tuples. Suppose the schema has the form

$$(\breve{\chi}_n \bullet \bigoplus_{m\in\mathbb{N}_0} \mu_n^{(m)} : H_n \times X \to \mathcal{P}(\coprod_{m\in\mathbb{N}_0} Y^m))_{n\in\mathbb{N}_0},$$

where $(\chi_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0))_{n\in\mathbb{N}_0}$ is a schema and $(\mu_n^{(m)} : H_n \times X \to \mathcal{P}(Y^m))_{n\in\mathbb{N}_0}$, for all $m \in \mathbb{N}_0$, is a schema for $\mathbf{y}^{(m)}$ given $\mathbf{x}$. It is assumed that $(\breve{\chi}_n \bullet \bigoplus_{m\in\mathbb{N}_0} \mu_n^{(m)})_{n\in\mathbb{N}_0}$ is constructed according to Proposition 3.2.3. Suppose that $Y = \prod_{p=1}^5 Y_p$, for some $Y_1, \ldots, Y_5$ and that each Y_p is a 'simple' space.

Deconstructing $(\breve{\chi}_n \bullet \bigoplus_{m\in\mathbb{N}_0} \mu_n^{(m)})_{n\in\mathbb{N}_0}$ simply reverses the construction of Proposition 3.2.3 and hence produces schemas having components with the following signatures.

$$\begin{aligned}
&\chi_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0)\\
&\mu_n^{(0)} : H_n \times X \to \mathcal{P}(\{()\})\\
&\mu_n^{(1)} : H_n \times X \to \mathcal{P}(Y)\\
&\qquad\vdots\\
&\mu_n^{(m)} : H_n \times X \to \mathcal{P}(Y^m)\\
&\qquad\vdots
\end{aligned}$$

The next step is to deconstruct $\mu^{(m)}$, for all $m \in \mathbb{N}$. Note that $\mathbf{y}_n^{(m)} : \Omega \to Y^m$, for all $m \in \mathbb{N}$. Thus $\mathbf{y}_n^{(m)} = (\mathbf{y}_n^{(m,1)}, \ldots, \mathbf{y}_n^{(m,m)})$, where $\mathbf{y}_n^{(m,j)} : \Omega \to Y$, for $j = 1, \ldots, m$. Suppose that the following conditional independence conditions are satisfied: for $j = 2, \ldots, m$,

$$\sigma(\mathbf{y}_n^{(m,j)}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,j-1)})}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,1)}, \ldots, \mathbf{y}_n^{(m,j-1)}),$$

for all $n \in \mathbb{N}_0$. (Other assumptions are possible, of course, in which case the following description would need to be modified accordingly.) Intuitively, these conditions state that each element (except the first element) of a list is conditionally independent of its predecessors given its immediate predecessor. This is illustrated in Figure 3.22.

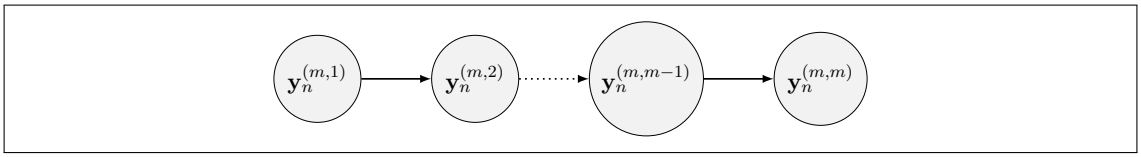

Figure 3.22: Fragment corresponding to a list of length m in the dependency graph

Then, under these conditional independence conditions,

$$\mu_n^{(m)} = \bigotimes_{j=1}^{m} \lambda(h, x, y_1, \ldots, y_{j-1}).\mu_n^{(m,j)}(h, x, y_{j-1}),$$

where

$$\begin{aligned}
&\mu_n^{(m,1)} : H_n \times X \to \mathcal{P}(Y) \\
&\mu_n^{(m,2)} : H_n \times X \times Y \to \mathcal{P}(Y) \\
&\qquad\qquad \vdots \\
&\mu_n^{(m,m)} : H_n \times X \times Y \to \mathcal{P}(Y),
\end{aligned}$$

for all $n \in \mathbb{N}_0$.

The final step is to deconstruct $\mu^{(m,j)}$, for $j = 1, \ldots, m$. Note that, since $\mathbf{y}_n^{(m,j)} : \Omega \to \prod_{p=1}^{5} Y_p$, then $\mathbf{y}_n^{(m,j)} = (\mathbf{y}_n^{(m,j,1)}, \ldots, \mathbf{y}_n^{(m,j,5)})$, where $\mathbf{y}_n^{(m,j,p)} : \Omega \to Y_p$, for $p = 1, \ldots, 5$.

First, $\mu^{(m,1)}$ is deconstructed. Suppose that the following conditional independence conditions are satisfied:

$$\begin{aligned}
&\sigma(\mathbf{y}_n^{(m,1,2)}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,1,1)}) \\
&\sigma(\mathbf{y}_n^{(m,1,4)}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,1,3)})}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,1,1)}, \mathbf{y}_n^{(m,1,2)}, \mathbf{y}_n^{(m,1,3)}) \\
&\sigma(\mathbf{y}_n^{(m,1,5)}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,1,3)})}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,1,1)}, \mathbf{y}_n^{(m,1,2)}, \mathbf{y}_n^{(m,1,3)}, \mathbf{y}_n^{(m,1,4)})
\end{aligned}$$

for all $n \in \mathbb{N}_0$. These conditions are illustrated by Figure 3.23.

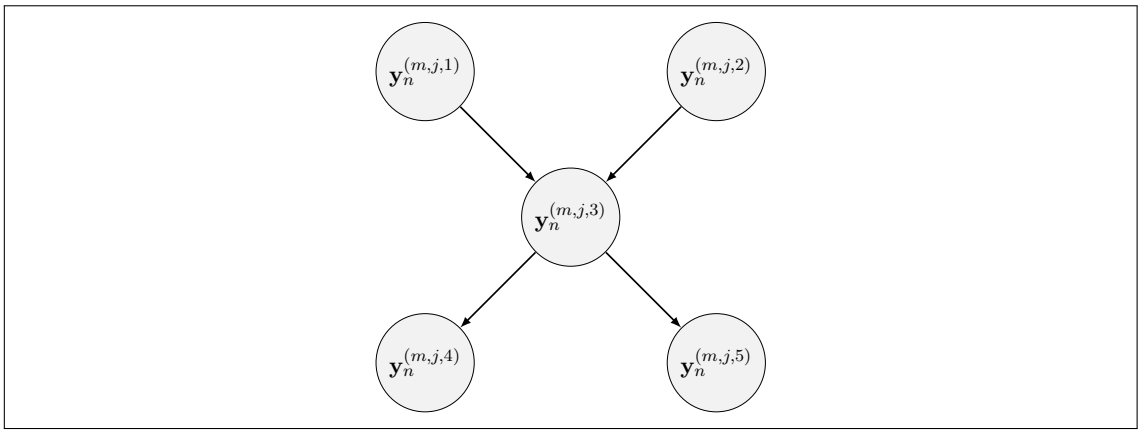

Figure 3.23: Fragment corresponding to a 5-tuple in the dependency graph

Then, under these conditional independence conditions,

$$\begin{aligned}
&\mu_n^{(m,1)} = \\
&\quad \mu_n^{(m,1,1)} \otimes (\lambda(h, x, y_1).\mu_n^{(m,1,2)}(h, x)) \otimes \mu_n^{(m,1,3)} \otimes \\
&\quad (\lambda(h, x, y_1, y_2, y_3).\mu_n^{(m,1,4)}(h, x, y_3)) \otimes (\lambda(h, x, y_1, y_2, y_3, y_4).\mu_n^{(m,1,5)}(h, x, y_3)),
\end{aligned}$$

where

$$\mu_n^{(m,1,1)} : H_n \times X \to \mathcal{P}(Y_1)$$
$$\mu_n^{(m,1,2)} : H_n \times X \to \mathcal{P}(Y_2)$$
$$\mu_n^{(m,1,3)} : H_n \times X \times Y_1 \times Y_2 \to \mathcal{P}(Y_3)$$
$$\mu_n^{(m,1,4)} : H_n \times X \times Y_3 \to \mathcal{P}(Y_4)$$
$$\mu_n^{(m,1,5)} : H_n \times X \times Y_3 \to \mathcal{P}(Y_5).$$

Now $\mu^{(m,j)}$, for $j = 2, \ldots, m$ is deconstructed. For $j = 2, \ldots, m$, suppose that the following conditional independence conditions are satisfied:

$$\sigma(\mathbf{y}_n^{(m,j,2)}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,j-1)})}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,j-1)}, \mathbf{y}_n^{(m,j,1)})$$
$$\sigma(\mathbf{y}_n^{(m,j,4)}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,j-1)}, \mathbf{y}_n^{(m,j,3)})}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,j-1)}, \mathbf{y}_n^{(m,j,1)}, \mathbf{y}_n^{(m,j,2)}, \mathbf{y}_n^{(m,j,3)})$$
$$\sigma(\mathbf{y}_n^{(m,j,5)}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,j-1)}, \mathbf{y}_n^{(m,j,3)})}{\perp\!\!\!\perp} \sigma(\mathbf{h}_n, \mathbf{x}_n, \mathbf{y}_n^{(m,j-1)}, \mathbf{y}_n^{(m,j,1)}, \mathbf{y}_n^{(m,j,2)}, \mathbf{y}_n^{(m,j,3)}, \mathbf{y}_n^{(m,j,4)})$$

for all $n \in \mathbb{N}_0$. These conditions are illustrated by Figure 3.23.

Then, under these conditional independence conditions,

$$\begin{aligned}\mu_n^{(m,j)} = \\ &\mu_n^{(m,j,1)} \otimes (\mu_n^{(m,j,2)} \circ \lambda(h,x,y,y_1).(h,x,y)) \otimes \mu_n^{(m,j,3)} \otimes \\ &(\mu_n^{(m,j,4)} \circ \lambda(h,x,y,y_1,y_2,y_3).(h,x,y,y_3)) \otimes \\ &\qquad\qquad (\mu_n^{(m,j,5)} \circ \lambda(h,x,y,y_1,y_2,y_3,y_4).(h,x,y,y_3)),\end{aligned}$$

where

$$\mu_n^{(m,j,1)} : H_n \times X \times Y \to \mathcal{P}(Y_1)$$
$$\mu_n^{(m,j,2)} : H_n \times X \times Y \to \mathcal{P}(Y_2)$$
$$\mu_n^{(m,j,3)} : H_n \times X \times Y \times Y_1 \times Y_2 \to \mathcal{P}(Y_3)$$
$$\mu_n^{(m,j,4)} : H_n \times X \times Y \times Y_3 \to \mathcal{P}(Y_4)$$
$$\mu_n^{(m,j,5)} : H_n \times X \times Y \times Y_3 \to \mathcal{P}(Y_5).$$

This completes the deconstruction. The complete set of schemas produced by the deconstruction is thus χ, $\mu^{(0)}$, and $\mu^{(m,j,p)}$, for all $m \in \mathbb{N}$, $j = 1, \ldots, m$, and $p = 1, \ldots, 5$. At any time step, the empirical belief base will contain the empirical beliefs obtained from these schemas.

Example 3.3.2. By way of contrast with Example 3.3.1, consider a schema whose codomain is the set of all probability measures on sets of 5-tuples. Suppose the schema has the form

$$(\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^Y)_{n \in \mathbb{N}_0},$$

where $Y = \prod_{p=1}^{5} Y_p$, for some $Y_1, \ldots, Y_5$.

Suppose that Y is countably infinite. (In case Y is finite, the changes needed are obvious.) Let $(y_n)_{n\in\mathbb{N}}$ be an enumeration of Y. Deconstructing μ produces schemas having components with the following signatures.

$$\begin{aligned}
&\mu_n^{(y_1)} : H_n \times X \to \mathcal{P}(\mathbb{B}) \\
&\mu_n^{(y_2)} : H_n \times X \times \mathbb{B} \to \mathcal{P}(\mathbb{B}) \\
&\qquad\qquad \vdots \\
&\mu_n^{(y_m)} : H_n \times X \times \mathbb{B}^{m-1} \to \mathcal{P}(\mathbb{B}) \\
&\qquad\qquad \vdots
\end{aligned}$$

Note that the structure of Y does not affect the deconstruction; Y is merely used as an index set. Thus this example produces a very different deconstruction to Example 3.3.1, even though the change from lists in Example 3.3.1 to sets here may appear to be comparatively minor.

Here is a summary of the results of this section. Consider a schema $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ of interest in some application. Usually, Y has some structure, say, it is a product or a sum. In this case, Propositions 3.3.1, 3.3.2, 3.3.3, and 3.3.4 show that μ can be deconstructed into a set $\{\mu_j\}_{j\in J}$ of 'simpler' schemas from which μ can be reconstructed. Here J is finite or countably infinite and 'simpler' means the space which supports the probability measures in the codomain is a constituent of the product or sum Y. Propositions 3.3.1, 3.3.2, 3.3.3, and 3.3.4 (and the deconstruction for weighted sums) show that any empirical belief obtained from μ can be reconstructed from a set of empirical beliefs obtained from the μ_j, for all $j \in J$. This process can be iterated until μ is deconstructed into a set of schemas that cannot be deconstructed further (or for which there is nothing to be gained by deconstructing further). In all cases, the design strategy is to deconstruct the original schema into a number of 'simpler' schemas that can be more easily maintained and utilized. In some applications, the original schema μ could be deconstructed into hundreds or even thousands of such schemas. Then any empirical belief obtained from μ can reconstructed from empirical beliefs obtained from the final set of schemas.

In the case of quotients, the situation is somewhat different. Having passed to the quotient schema, the original schema (generally) cannot be reconstructed exactly. However, the advantage may be that the quotient schema is 'simpler' than the original schema and 'enough' of the original schema can be reconstructed for the needs of the application. This aspect of quotients is further explained and illustrated below. Proposition 3.3.5 shows that an empirical belief obtained from the quotient schema is a quotient of the corresponding empirical belief obtained from the original schema.

Thus, after deconstruction, a typical empirical belief has a signature of the form $X \to \mathcal{P}(Y)$, where X structured but Y is not. The empirical belief may be piecewise-constant. Also, typically, Y is $\mathbb{R}$ or $\mathbb{R}^k$, for some k, and the distributions of interest are Gaussian distributions on $\mathbb{R}$ or $\mathbb{R}^k$, Y is $\mathbb{N}_0$ and the distributions of interest are Poisson distributions on $\mathbb{N}_0$, or Y is a finite space and the distributions of interest are categorical distributions on Y.

3.4 Representation Issues

So far, discussion of the structure of Y in a signature $X \to \mathcal{P}(Y)$ has been expressed in the mathematical language of product, sum, and quotient. To make the meaning of these mathematical terms clearer, they are now related to the more familiar language of the data types of belief representation. (Since the emphasis here is on beliefs instead of knowledge, the term 'belief representation' is often used in the following instead of 'knowledge representation'.) The data types of interest include tuples, lists, strings, sequences, sets, multisets, graphs, and function spaces. There are more data types just mentioned than there are corresponding mathematical concepts because some data types coincide mathematically but are thought of in different ways as data types, for example, by supporting a different set of functions that operate on the data type.

Consider first the case of tuples. The set of all tuples of the form $(x_1, \ldots, x_m)$, where $x_i \in X_i$, for $i = 1, \ldots, m$, is $\prod_{i=1}^{m} X_i$, a finite product space.

Consider the case of sets. The set of all subsets of some (countable) set X is $\mathbb{B}^X$. Here, $\mathbb{B}^X$ is the set of all $\mathbb{B}$-valued functions (that is, predicates) on X. Equivalently, one can think of $\mathbb{B}^X$ as the product space of $|X|$ copies of $\mathbb{B}$.

The next case is that of multisets. The set of all multisets based on some (countable) set X is $\mathbb{N}_0^X$. Here, $\mathbb{N}_0^X$ is the set of all $\mathbb{N}_0$-valued functions on X. Equivalently, one can think of $\mathbb{N}_0^X$ as the product space of $|X|$ copies of $\mathbb{N}_0$.

Now consider the case of (finite) lists. The set of all lists whose members are elements of some set X is $\coprod_{m \in \mathbb{N}_0} X^m$. Here, X^m can be thought of as the set of all lists of length m whose elements are in X. X^0 is $\{()\}$, the set consisting of the empty tuple which can be thought of as the empty list. Forming the sum over all $m \in \mathbb{N}_0$, then gives the set of all lists whose elements are in X. (In this particular case, the sum is the same as the union.)

This next case is that of strings. Let A be an alphabet. Then the set of finite strings over A is $\coprod_{n \in \mathbb{N}_0} A^n$. Thus, mathematically, (finite) strings and lists are the same, but as data types they have slightly different meanings because, for strings, A is a set of 'symbols', but for a list, A can be completely arbitrary. Also each data type supports a different set of functions that operate on them. Furthermore, different notation is used in each case: a list is usually denoted $[a_1, \ldots, a_n]$, but the corresponding string is denoted $a_1 \ldots a_n$. The set of finite *bit* strings is $\coprod_{n \in \mathbb{N}_0} \mathbb{B}^n$.

The set of all infinite strings over A is $A^{\mathbb{N}}$. Thus the set of all infinite strings is a product space. The set of infinite *bit* strings is $\mathbb{B}^{\mathbb{N}}$.

The set of all finite sequences whose members are elements of some set X is $\coprod_{n \in \mathbb{N}_0} X^n$. Thus finite sequences and (finite) lists are mathematically identical; in fact, 'finite sequence' and 'list' have the same intuitive meaning.

The set of all infinite sequences whose members are elements of some set X is $X^{\mathbb{N}}$. Thus the set of all infinite sequences is a product space where the exponent is $\mathbb{N}$.

The next case is graphs. The set of all directed graphs whose vertices are chosen from a (countable) set V is

$$\{(\nu, \varepsilon) \in \mathbb{B}^V \times \mathbb{B}^{V \times V} \mid \text{for all } v, w \in V,\ \varepsilon(v, w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}.$$

Thus the set of directed graphs is a subset of $\mathbb{B}^V \times \mathbb{B}^{V \times V}$. The edge (v, w) goes from vertex v to vertex w.

For undirected graphs, some notation is needed. Let X be a set. Let

$$X^{(2)} \triangleq \{\rho : X \to \{0,1\} \mid \sum_{x \in X} \rho(x) = 2\}.$$

One can identify $X^{(2)}$ with the set of subsets of X containing precisely two (distinct) elements. That is, $X^{(2)}$ can be identified with $\{\{x,y\} \mid x,y \in X, x \neq y\}$.

Then the set of all undirected graphs (without self-loops) whose vertices are chosen from a (countable) set V is

$$\{(\nu,\varepsilon) \in \mathbb{B}^V \times \mathbb{B}^{V^{(2)}} \mid \text{for all } \{v,w\} \in V^{(2)},\ \varepsilon(\{v,w\}) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}.$$

Thus the set of undirected graphs is a subset of $\mathbb{B}^V \times \mathbb{B}^{V^{(2)}}$. The edge $\{v,w\}$ joins vertex v and vertex w.

There are, of course, also other special kinds of graphs such as trees that are not discussed here.

The final case is function spaces. These have the form Y^X, which is the set of all functions from X to Y. Note that Y^X can also be thought of as a product space, in fact, the product of $|X|$ copies of Y. Function spaces include sets and multisets as special cases; for sets, Y is $\mathbb{B}$ and, for multisets, Y is $\mathbb{N}_0$.

In summary, this discussion shows that products (which include function spaces as special cases) and sums are sufficent to (mathematically) define all the data types mentioned. In the subsequent development of schemas and empirical beliefs, products and sums will be emphasized and the corresponding data type only mentioned to provide some intuition for intended meaning of the mathematical object.

Here is some more detail on the deconstruction of schemas and empirical beliefs when the codomain is specialized to the cases of probability measures on sets, multisets, lists, graphs, quotients, and function spaces.

3.4.1 Sets

In this case, there is some countable set Z such that schemas have the form $\mu = (\mu_n)_{n \in \mathbb{N}_0}$, where

$$\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^Z),$$

for all $n \in \mathbb{N}_0$. In the following discussion, Z is assumed to be countably infinite. (If Z is finite, the changes needed are obvious.) Thus Z is isomorphic to $\mathbb{N}$ and has the form $(z_m)_{m \in \mathbb{N}}$, where it is assumed that this sequence also provides the topological order of the corresponding vertices in the dependency graph. The deconstruction of $\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^Z)$, and corresponding empirical beliefs, are given in Figure 3.24 using Proposition 3.3.3.

The deconstruction in Figure 3.24 may be of benefit since the codomain $\mathcal{P}(\mathbb{B}^Z)$ of the original schema has been replaced by simpler codomains of the form $\mathcal{P}(\mathbb{B})$ in the constituents, which can be easier to acquire and utilize. However, a remaining issue that has to be dealt with is that Z is usually infinite. Towards this, note that the main reason for wanting Z to be infinite is not to allow the consideration of infinite subsets of Z but instead *finite* subsets of Z of *unbounded* size. Thus attention is mainly restricted to $\mathcal{F}_Z$,

$$\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^Z)$$
$$\Downarrow$$
$$\begin{cases} \mu_n^{(1)} : H_n \times X \to \mathcal{P}(\mathbb{B}) \\ \mu_n^{(2)} : H_n \times X \times \mathbb{B} \to \mathcal{P}(\mathbb{B}) \\ \quad\vdots \\ \mu_n^{(m)} : H_n \times X \times \mathbb{B}^{m-1} \to \mathcal{P}(\mathbb{B}) \\ \quad\vdots \end{cases}$$

$$\lambda x.\mu_n(h,x) : X \to \mathcal{P}(\mathbb{B}^Z)$$
$$\Downarrow$$
$$\begin{cases} \lambda x.\mu_n^{(1)}(h,x) : X \to \mathcal{P}(\mathbb{B}) \\ \lambda(x,b_1).\mu_n^{(2)}(h,x,b_1) : X \times \mathbb{B} \to \mathcal{P}(\mathbb{B}) \\ \quad\vdots \\ \lambda(x,b_1,\ldots,b_{m-1}).\mu_n^{(m)}(h,x,b_1,\ldots,b_{m-1}) : X \times \mathbb{B}^{m-1} \to \mathcal{P}(\mathbb{B}) \\ \quad\vdots \end{cases}$$

Figure 3.24: Deconstruction of schemas and empirical beliefs in the case of sets

the set of *finite* subsets of Z. In this case, progress can be made, and this is the case that is most useful for applications. So the following discussion is limited to probability measures on $\mathbb{B}^Z$ whose support is contained in $\mathcal{F}_Z$.

Let $F \triangleq \{z_1, \ldots, z_M\}$ be a (non-empty) finite subset of Z. (How F is chosen is discussed below.) For $k = 1, \ldots, M$, let

$$\mu_n^{(k)} : H_n \times X \times \mathbb{B}^{k-1} \to \mathcal{P}(\mathbb{B})$$

be a schema component. For $k > M$, define the schema component

$$\mu_n^{(k)} : H_n \times X \times \mathbb{B}^{k-1} \to \mathcal{P}(\mathbb{B})$$

by

$$\mu_n^{(k)}(h, x, b_1, \ldots, b_{k-1}) = \delta_{\mathsf{F}},$$

for all $h \in H_n$, $x \in X$, $(b_1, \ldots, b_{k-1}) \in \mathbb{B}^{k-1}$. (Here, δ_{F} is the Dirac measure at F.) Now put $\mu_n = \bigotimes_{m \in \mathbb{N}} \mu_n^{(m)}$, for all $n \in \mathbb{N}_0$. Then, by Proposition A.8.9, μ has the property that

$$\mu_n(h,x)(\{s \in \mathbb{B}^Z \mid s(z) = \mathsf{F}, \text{ for all } z \in Z \setminus F\}) = 1,$$

for all $h \in H_n$, $x \in X$, and $n \in \mathbb{N}_0$. Thus, using the probability measure $\mu_n(h,x)$, the probability of the set of subsets of F is 1.

By the way, the schema

$$\mu_n^{(k)} : H_n \times X \times \mathbb{B}^{k-1} \to \mathcal{P}(\mathbb{B})$$

could also be denoted

$$\mu_n^{(k)} : H_n \times X \times \mathbb{B}^{\{z_1,\ldots,z_{k-1}\}} \to \mathcal{P}(\mathbb{B}^{\{z_k\}}),$$

since $\mathbb{B}^{\{z_1,\ldots,z_{k-1}\}}$ is isomorphic to $\mathbb{B}^{k-1}$ and $\mathbb{B}^{\{z_k\}}$ is isomorphic to $\mathbb{B}$. The latter notation for the schema is more precise and makes clear how the schema is associated with the elements $z_1, \ldots, z_k$ of Z; in most practical applications, knowing exactly which elements of Z are in F is important. The former notation, which is less intrusive and used throughout, should be interpreted with the meaning of the latter notation.

The probability kernels $\mu_n^{(k)}$, for $k > M$, are Dirac kernels which guarantee that subsets sampled from the distribution in the codomain of μ_n are subsets of F. The more interesting probability kernels are the $\mu_n^{(k)}$, for $k \in \{1, \ldots, M\}$. These are probability kernels that are acquired by filtering. Assuming that it is appropriate to model these probability kernels with noisy-OR probability kernels, the number of parameters that need to be learned is linear in the number of arguments (precisely, the dimension of the domain plus one). Without some kind of linearity assumption like this, the size of F would have to be considerably constrained for the model to be tractable. In fact, [113, Section 26.5.4] describes a Bayesian network with millions of vertices, where only noisy-OR probability kernels are used, for which both the parameters of the probability kernels and structure of the graph were learned from data. Thus the much smaller models envisioned here for probability measures on the subsets of a finite set having, say, a few thousand elements are computationally feasible. The other possibility for ensuring tractability is that strong conditional independence assumptions imply that each vertex has only few parents in the dependency graph. Note that sampling from a distribution on $\mathbb{B}^Z$ can be done by ancestral sampling from the product distribution on the part indexed by elements of F and then adding F's for the remaining elements of the set.

Now the discussion turns to the choice of F. For the intended applications, any empirical belief maintained by the agent will be changing over time as new observations are received. These changes are partly manifested in changes to F. This aspect is quite different to the application described in [113, Section 26.5.4] and means that both the parameters of each constituent and the graph structure of the probability measure on the subsets will need to be learned on-line. At the least, F must contain all the elements that appear in (relevant) sets that appear in observations received so far. When an element of Z that is not in F appears in a set in an observation, then F is extended by this element. The new graph structure and the parameters of the constituents then need to be learned from the data points received so far. In most cases, it would be expected that much of the previous model would survive and so only incremental changes that can be carried out efficiently would be needed. However, from time to time, extending F will involve building a significantly different model and thus will be computationally expensive. This aspect provides the greatest constraint on how large F can be.

An intuitively obvious property of the schema $\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)}$ constructed above is that an integral of the form $\int_{\mathbb{B}^Z} f \; d(\bigotimes_{m\in\mathbb{N}} \mu_n^{(m)})(h, x)$ can be reduced to an integral over the space $\mathbb{B}^F$. The technical details of this are given by Proposition A.8.9. In effect, f' in

that proposition is f restricted to the subsets of F. The integral $\int_{\mathbb{B}^F} f' \, d(\bigotimes_{j=1}^{M} \mu_n^{(j)})(h,x)$ can be computed efficiently by Monte Carlo methods. Proposition A.8.9 shows that representing finite sets and integrating functions over such sets is not much different to dealing with conventional Bayesian networks for which the vertices are boolean-valued. The major difference is the on-line nature of the intended applications considered here that implies the need to modify the graph structure and constituent parameters as more observations are received.

Example 3.4.1. Consider states that are subsets of some set Z. Thus the state schema has components with signature

$$\mu_n : H_n \to \mathcal{P}(\mathbb{B}^Z).$$

The transition model has components with signature

$$\tau_n : A \times \mathbb{B}^Z \to \mathcal{P}(\mathbb{B}^Z).$$

The observation model has components with signature

$$\xi_n : \mathbb{B}^Z \to \mathcal{P}(O).$$

Note that the definitions of τ_n and ξ_n will require predicates whose domain is $\mathbb{B}^Z$, since $\mathbb{B}^Z$ is a factor in their domains. The filtering results of Chapter 2 can be applied to this example at the level of the undeconstructed state schema, although if Z is infinite, some care will need to be taken to effectively work with a finite subset. Alternatively, under appropriate conditional independence assumptions, the schema could be deconstructed according to Figure 3.24 (modified to take the conditional independence assumptions into account) and the more general filtering methods of Chapter 4 applied.

The case of sets is now considered more generally. Let $F \triangleq \{z_1, \ldots, z_M\}$ be a (non-empty) finite subset of Z. Consider the equivalence relation $\sim$ on Z that gives the partition $Z/\!\sim = \{\{z_1\}, \ldots, \{z_M\}, Z \setminus F\}$ of Z. Let $\pi : Z \to Z/\!\sim$ be the canonical surjection. Define $p : \mathbb{B}^{Z/\sim} \to \mathbb{B}^Z$ by $p(f) = f \circ \pi$, for all $f \in \mathbb{B}^{Z/\sim}$. Let $(\nu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^{Z/\sim}))_{n\in\mathbb{N}_0}$ be a schema with the property that $\nu_n(h,x)(\{s : \mathbb{B}^{Z/\sim} \mid s(Z \setminus F) = \mathsf{F}\}) = 1$, for all $x \in X$ and $h \in H_n$. Then the schemas considered so far have the form $(\lambda(h,x).(\nu_n(h,x) \circ p^{-1}))_{n\in\mathbb{N}_0}$, since

$$\begin{aligned}
&\{s' \in \mathbb{B}^{Z/\sim} \mid s'(Z \setminus F) = \mathsf{F}\} \\
&= p^{-1}(\{s' \circ \pi \in \mathbb{B}^Z \mid s'(Z \setminus F) = \mathsf{F}\}) \\
&= p^{-1}(\{s \in \mathbb{B}^Z \mid s(z) = \mathsf{F}, \text{ for all } z \in Z \setminus F\}),
\end{aligned}$$

and so

$$\begin{aligned}
&(\nu_n(h,x) \circ p^{-1})(\{s \in \mathbb{B}^Z \mid s(z) = \mathsf{F}, \text{ for all } z \in Z \setminus F\}) \\
&= \nu_n(h,x)(\{s' \in \mathbb{B}^{Z/\sim} \mid s'(Z \setminus F) = \mathsf{F}\}) \\
&= 1,
\end{aligned}$$

for all $h \in H_n$ and $x \in X$. This suggests the following generalization.

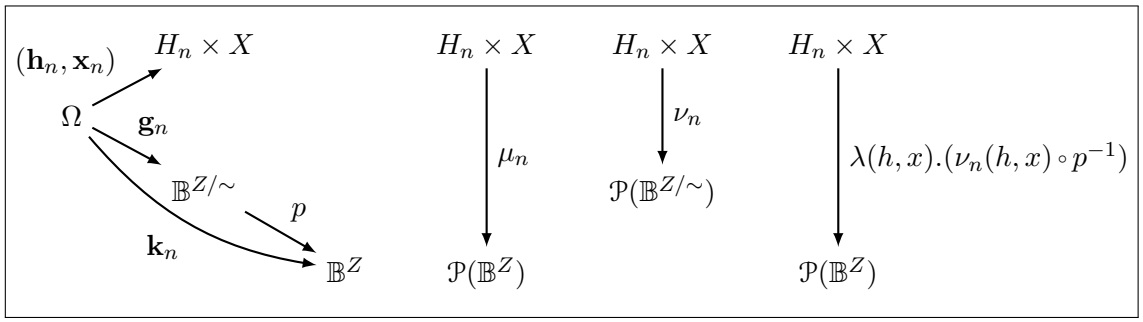

Figure 3.25: Setting for sets

The generalized setting for sets is illustrated in Figure 3.25. There are stochastic processes $\mathbf{k} : \Omega \to (\mathbb{B}^Z)^{\mathbb{N}_0}$ and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$, and a probability kernel $\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^Z)$ that is a regular conditional distribution of $\mathbf{k}_n$ given $(\mathbf{h}_n, \mathbf{x}_n)$, for all $n \in \mathbb{N}_0$. The task is to acquire $(\mu_n)_{n\in\mathbb{N}_0}$ (or at least some close approximation of it).

Let $\sim$ be an equivalence relation on Z such that the set of equivalence classes $Z/\sim$ is finite. Let $\pi : Z \to Z/\sim$ be the canonical surjection. Define $p : \mathbb{B}^{Z/\sim} \to \mathbb{B}^Z$ by $p(f) = f \circ \pi$, for all $f \in \mathbb{B}^{Z/\sim}$. By Part 1 of Proposition A.10.5, p is measurable. Let $(\nu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^{Z/\sim}))_{n\in\mathbb{N}_0}$ be the schema for $\mathbf{g}$, say, given $(\mathbf{h}, \mathbf{x})$. By Parts 3 and 4 of Proposition A.10.5, $(\lambda(h,x).(\nu_n(h,x) \circ p^{-1}) : H_n \times X \to \mathcal{P}(\mathbb{B}^Z))_{n\in\mathbb{N}_0}$ is the schema for $\lambda\omega.((p \circ \mathbf{g}_n)(\omega))_{n\in\mathbb{N}_0}$ given $(\mathbf{h}, \mathbf{x})$. Thus, for all $n \in \mathbb{N}_0$, $h \in H_n$, and $x \in X$, $\nu_n(h,x) \circ p^{-1} : \mathcal{P}(\mathbb{B}^Z)$ is an empirical belief.

If, for all $h \in H_n$ and $x \in X$, $f : \mathbb{B}^Z \to \mathbb{R}$ is a $(\nu_n(h,x) \circ p^{-1})$-integrable function, then Part 6 of Proposition A.10.5 shows that

$$\int_{\mathbb{B}^Z} f \; d(\nu_n(h,x) \circ p^{-1}) = \int_{\mathbb{B}^{Z/\sim}} f \circ p \; d\nu_n(h,x).$$

The latter integral can be easily computed by Monte Carlo methods since $\mathbb{B}^{Z/\sim}$ is a finite product space.

Let C be the set of functions in $\mathbb{B}^Z$ that are constant on each equivalence class in the partition of Z. (As indicated above, by a suitable choice of equivalence relation, C includes all the predicates corresponding to subsets of Z up to any desired finite size.) Then Part 5 of Proposition A.10.5 shows that $(\nu_n(h,x) \circ p^{-1})(C) = 1$, for all $h \in H_n$ and $x \in X$. Thus the support of the probability measure $\nu_n(h,x) \circ p^{-1}$ is the set of piecewise-constant functions in $\mathbb{B}^Z$ that are based on the partition induced by $\sim$. The class C provides a useful space of sets provided that the equivalence relation has been carefully chosen. How $\sim$ is chosen and $\nu_n(h_n)$ is learned is discussed in Chapter 4.

Further detail is given about this case in the more general case of function spaces is given later in this section.

3.4.2 Multisets

In this case, there is some countable set Z such that schemas have the form $(\mu_n)_{n\in\mathbb{N}_0}$, where

$$\mu_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0^Z),$$

for all $n \in \mathbb{N}_0$. In the following discussion, Z is assumed to be countably infinite. Thus Z is isomorphic to $\mathbb{N}$ and has the form $(z_m)_{m\in\mathbb{N}}$, where it is assumed that this sequence also provides the topological order of the corresponding vertices in the dependency graph. The deconstruction of $\mu_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0^Z)$, and corresponding empirical beliefs, is given in Figure 3.26 using Proposition 3.3.3.

$$
\begin{array}{l}
\mu_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0^Z) \\
\qquad\qquad \Downarrow \\
\left\{ \begin{array}{l}
\mu_n^{(1)} : H_n \times X \to \mathcal{P}(\mathbb{N}_0) \\
\mu_n^{(2)} : H_n \times X \times \mathbb{N}_0 \to \mathcal{P}(\mathbb{N}_0) \\
\qquad\qquad \vdots \\
\mu_n^{(m)} : H_n \times X \times \mathbb{N}_0^{m-1} \to \mathcal{P}(\mathbb{N}_0) \\
\qquad\qquad \vdots
\end{array} \right.
\end{array}
$$

$$
\begin{array}{l}
\lambda x.\mu_n(h,x) : X \to \mathcal{P}(\mathbb{N}_0^Z) \\
\qquad\qquad \Downarrow \\
\left\{ \begin{array}{l}
\lambda x.\mu_n^{(1)}(h,x) : X \to \mathcal{P}(\mathbb{N}_0) \\
\lambda (x,n_1).\mu_n^{(2)}(h,x,n_1) : X \times \mathbb{N}_0 \to \mathcal{P}(\mathbb{N}_0) \\
\qquad\qquad \vdots \\
\lambda (x,n_1,\ldots,n_{m-1}).\mu_n^{(m)}(h,x,n_1,\ldots,n_{m-1}) : X \times \mathbb{N}_0^{m-1} \to \mathcal{P}(\mathbb{N}_0) \\
\qquad\qquad \vdots
\end{array} \right.
\end{array}
$$

Figure 3.26: Deconstruction in the case of multisets

Once again, attention is concentrated of the set of finite multisets, that is, those $m \in \mathbb{N}_0^Z$ for which there exists a finite subset F of Z (dependent on m) such that $m(x) = 0$, for all $x \in Z \setminus F$. The extra complication in the multiset case compared to the set case is that $\mathbb{N}_0$ is infinite. In practice, there are typically small numbers of duplicates of an element in a multiset so this doesn't cause too much extra complication, but there is necessarily more computation needed to handle probability measures on multisets than for sets.

Let $F \triangleq \{z_1, \ldots, z_M\}$ be a (non-empty) finite subset of Z. For $k = 1, \ldots, M$, let

$$\mu_n^{(k)} : H_n \times X \times \mathbb{N}_0^{k-1} \to \mathcal{P}(\mathbb{N}_0)$$

be a schema component. For $k > M$, define the probability kernel

$$\mu_n^{(k)} : H_n \times X \times \mathbb{N}_0^{k-1} \to \mathcal{P}(\mathbb{N}_0)$$

by

$$\mu_n^{(k)}(h, x, m_1, \ldots, m_{k-1}) = \delta_0,$$

for all $h \in H_n$, $x \in X$, $(m_1, \ldots, m_{k-1}) \in \mathbb{N}_0^{k-1}$. Now put $\mu_n = \bigotimes_{m \in \mathbb{N}} \mu_n^{(m)}$, for all $n \in \mathbb{N}_0$. Then, by Proposition A.8.9, each μ_n has the property that

$$\mu_n(h,x)(\{m \in \mathbb{N}_0^Z \mid m(z) = 0, \text{ for all } z \in Z \setminus F\}) = 1,$$

for all $h \in H_n$ and $x \in X$. Thus, using the probability measure $\mu_n(h,x)$, the probability of the set of multisets whose support is contained in F is 1.

The schema

$$\mu_n^{(k)} : H_n \times X \times \mathbb{N}_0^{k-1} \to \mathcal{P}(\mathbb{N}_0)$$

could also be denoted

$$\mu_n^{(k)} : H_n \times X \times \mathbb{N}_0^{\{z_1,\ldots,z_{k-1}\}} \to \mathcal{P}(\mathbb{N}_0^{\{z_k\}}),$$

since $\mathbb{N}_0^{\{z_1,\ldots,z_{k-1}\}}$ is isomorphic to $\mathbb{N}_0^{k-1}$ and $\mathbb{N}_0^{\{z_k\}}$ is isomorphic to $\mathbb{N}_0$. The latter notation for the schema is more precise and makes clear how the schema is associated with the elements $z_1, \ldots, z_k$ of Z. Throughout, the former notation should be interpreted with the meaning of the latter notation.

At the least, F must contain all the elements that appear in (relevant) multisets that appear in observations received so far. When an element of Z that is not in F appears in a set in an observation, then F is extended by this element. The new graph structure and the parameters of the constituents then need to be learned from the data points received so far.

Similarly to sets, the schema $\bigotimes_{m \in \mathbb{N}} \mu_n^{(m)}$ has the property that an integral of the form $\int_{\mathbb{N}_0^Z} f \; d(\bigotimes_{m \in \mathbb{N}} \mu_n^{(m)})(h,x)$ can be reduced to an integral over the space $\mathbb{N}_0^F$. The technical details of this are given by Proposition A.8.10. In effect, f' in that proposition is f restricted to the multisets whose support is contained in F. The integral $\int_{\mathbb{N}_0^F} f' \, d(\bigotimes_{j=1}^m \mu_n^{(j)})(h,x)$ can be computed efficiently by Monte Carlo methods.

In an analogous way to sets, the above approach can be generalized. This generalized setting for multisets is illustrated in Figure 3.27. There are stochastic processes $\mathbf{k} : \Omega \to (\mathbb{N}_0^Z)^{\mathbb{N}_0}$ and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$, and a probability kernel $\mu_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0^Z)$ that is a regular conditional distribution of $\mathbf{k}_n$ given $(\mathbf{h}_n, \mathbf{x}_n)$, for all $n \in \mathbb{N}_0$. The task is to acquire $(\mu_n)_{n \in \mathbb{N}_0}$ (or at least some close approximation of it).

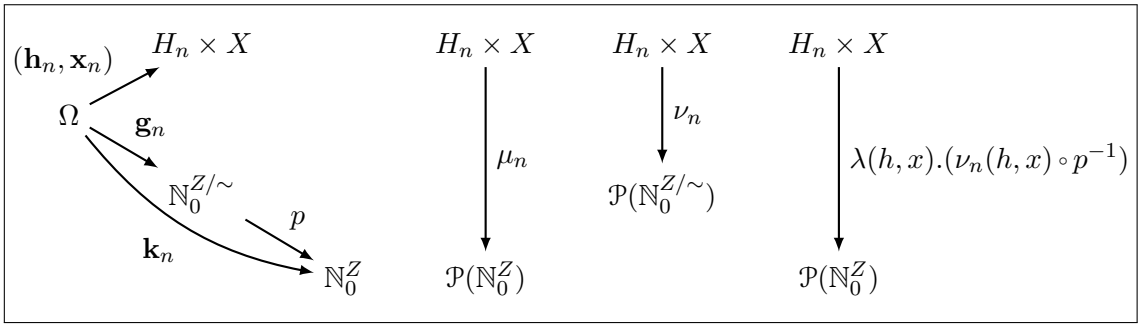

Figure 3.27: Setting for multisets

Let $\sim$ be an equivalence relation on Z such that the set of equivalence classes $Z/\sim$ is finite. Let $\pi : Z \to Z/\sim$ be the canonical surjection. Define $p : \mathbb{N}_0^{Z/\sim} \to \mathbb{N}_0^Z$ by $p(f) = f \circ \pi$, for all $f \in \mathbb{N}_0^{Z/\sim}$. By Part 1 of Proposition A.10.5, p is measurable. Let $(\nu_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0^{Z/\sim}))_{n\in\mathbb{N}_0}$ be the schema for $\mathbf{g}$, say, given $(\mathbf{h}, \mathbf{x})$. By Parts 3 and 4 of Proposition A.10.5, $(\lambda(h,x).(\nu_n(h,x) \circ p^{-1}) : H_n \times X \to \mathcal{P}(\mathbb{N}_0^Z))_{n\in\mathbb{N}_0}$ is the schema for $\lambda\omega.((p \circ \mathbf{g}_n)(\omega))_{n\in\mathbb{N}_0}$ given $(\mathbf{h}, \mathbf{x})$. Thus, for all $n \in \mathbb{N}_0$, $h \in H_n$, and $x \in X$, $\nu_n(h,x) \circ p^{-1} : \mathcal{P}(\mathbb{N}_0^Z)$ is an empirical belief.

If, for all $h \in H_n$ and $x \in X$, $f : \mathbb{N}_0^Z \to \mathbb{R}$ is a $(\nu_n(h,x) \circ p^{-1})$-integrable function, then Part 6 of Proposition A.10.5 shows that

$$\int_{\mathbb{N}_0^Z} f \; d(\nu_n(h,x) \circ p^{-1}) = \int_{\mathbb{N}_0^{Z/\sim}} f \circ p \; d\nu_n(h,x).$$

The latter integral can be easily computed by Monte Carlo methods since $\mathbb{N}_0^{Z/\sim}$ is a finite product space.

Let C be the set of functions in $\mathbb{N}_0^Z$ that are constant on each equivalence class in the partition of Z. Then Part 5 of Proposition A.10.5 shows that $(\nu_n(h,x) \circ p^{-1})(C) = 1$, for all $h \in H_n$ and $x \in X$. Thus the support of the probability measure $\nu_n(h,x) \circ p^{-1}$ is the set of piecewise-constant functions in $\mathbb{N}_0^Z$ that are based on the partition induced by $\sim$. The class C provides a useful space of multisets provided that the equivalence relation has been carefully chosen. How $\sim$ is chosen and $\nu_n(h_n)$ is learned is discussed in Chapter 4.

Further detail is given about this case in the more general case of function spaces is given later in this section.

3.4.3 Lists

Let $(Z, \mathcal{Z})$ be a measurable space. One can regard $\coprod_{m\in\mathbb{N}_0} Z^m$ as the set of all (finite) lists whose elements are in the set Z, where Z^m is the set of such lists of length m. Consider a schema of the form

$$(\breve{\chi}_n \bullet \bigoplus_{m\in\mathbb{N}_0} \mu_n^{(m)} : H_n \times X \to \mathcal{P}(\coprod_{m\in\mathbb{N}_0} Z^m))_{n\in\mathbb{N}_0},$$

for some $(\chi_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0))_{n\in\mathbb{N}_0}$ and $(\mu_n^{(m)} : H_n \times X \to \mathcal{P}(Z^m))_{n\in\mathbb{N}_0}$, for all $m \in \mathbb{N}_0$ (as in Proposition 3.2.3). This can be thought of as a schema whose codomain is distributions on lists. Figure 3.28 shows the deconstruction, given by the deconstruction for the case of weighted sums, for this schema.

In practice, one can restrict attention to just those lists up to some maximum length. To see this, recall that

$$(\breve{\chi}_n \bullet \bigoplus_{m\in\mathbb{N}_0} \mu_n^{(m)})(h,x)(\coprod_{m\in\mathbb{N}_0} A_m) = \sum_{m\in\mathbb{N}_0} \breve{\chi}_n(x,h)(m) \; \mu_n^{(m)}(h,x)(A_m),$$

for all $n \in \mathbb{N}_0$, $h \in H_n$, $x \in X$, and $\coprod_{m\in\mathbb{N}_0} A_m \in \bigoplus_{m\in\mathbb{N}_0} \mathcal{Z}^{\otimes m}$. (Here, $\mathcal{Z}^{\otimes m} \triangleq \bigotimes_{j=1}^m \mathcal{Z}j$, where $\mathcal{Z}_j \triangleq \mathcal{Z}$, for $j = 1, \ldots, m$.) Now, for all $n \in \mathbb{N}_0$, $h \in H_n$, $x \in X$, and $\varepsilon > 0$, there exists $K \in \mathbb{N}_0$ such that

$$\sum_{m>K} \breve{\chi}_n(h,x)(m) < \varepsilon.$$

Thus, for all $n \in \mathbb{N}_0$, $h \in H_n$, and $x \in X$,

$$\sum_{m>K} \breve{\chi}_n(h,x)(m)\ \mu_n^{(m)}(h,x)(A_m) < \epsilon.$$

This means that, for ε chosen to be sufficiently small, it is possible to ignore lists of length greater than K because they have a sufficiently low probability of existing. However, K depends upon n, so that it may be necessary to adjust the value of K over time. Note that, interpreted in the obvious way, this remark applies more generally to any weighted sum since it only depends on a property of $\breve{\chi}_n$.

$$\breve{\chi}_n \bullet \bigoplus_{m\in\mathbb{N}_0} \mu_n^{(m)} : H_n \times X \to \mathcal{P}(\coprod_{m\in\mathbb{N}_0} Z^m)$$

$$\Downarrow$$

$$\begin{cases} \chi_n : H_n \times X \to \mathcal{P}(\mathbb{N}_0) \\ \mu_n^{(0)} : H_n \times X \to \mathcal{P}(\{()\}) \\ \mu_n^{(1)} : H_n \times X \to \mathcal{P}(Z) \\ \qquad\vdots \\ \mu_n^{(m)} : H_n \times X \to \mathcal{P}(Z^m) \\ \qquad\vdots \end{cases}$$

$$\lambda x.\breve{\chi}_n(h,x) \bullet \bigoplus_{m\in\mathbb{N}_0} \lambda x.\mu_n^{(m)}(h,x) : X \to \mathcal{P}(\coprod_{m\in\mathbb{N}_0} Z^m)$$

$$\Downarrow$$

$$\begin{cases} \lambda x.\chi_n(h,x) : X \to \mathcal{P}(\mathbb{N}_0) \\ \lambda x.\mu_n^{(0)}(h,x) : X \to \mathcal{P}(\{()\}) \\ \lambda x.\mu_n^{(1)}(h,x) : X \to \mathcal{P}(Z) \\ \qquad\vdots \\ \lambda x.\mu_n^{(m)}(h,x) : X \to \mathcal{P}(Z^m) \\ \qquad\vdots \end{cases}$$

Figure 3.28: Deconstruction in the case of lists

3.4.4 Graphs

As noted earlier in this section, the set of all directed graphs whose vertices are chosen from a set V is

$$\{(\nu,\varepsilon) \in \mathbb{B}^V \times \mathbb{B}^{V\times V} \mid \text{for all } v,w \in V,\ \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\},$$

that is, directed graphs are a subset of $\mathbb{B}^V \times \mathbb{B}^{V\times V}$. Now consider the schema

$$(\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^V \times \mathbb{B}^{V\times V}))_{n\in\mathbb{N}_0}.$$

For the codomain of the schema to correctly model distributions over directed graphs, it will be necessary to ensure that

$$\mu_n(h,x)(\{(\nu,\varepsilon) \in \mathbb{B}^V \times \mathbb{B}^{V\times V} \mid \text{for all } v,w \in V, \\ \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}) = 1, \quad (3.4.1)$$

for all $h \in H_n$, $x \in X$, and $n \in \mathbb{N}_0$.

$$\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^V \times \mathbb{B}^{V\times V})$$
$$\Downarrow$$
$$\begin{cases} \mu_n^{(1)} : H_n \times X \to \mathcal{P}(\mathbb{B}^V) \\ \mu_n^{(2)} : H_n \times X \times \mathbb{B}^V \to \mathcal{P}(\mathbb{B}^{V\times V}) \end{cases}$$

$$\lambda x.\mu_n(h,x) : X \to \mathcal{P}(\mathbb{B}^V \times \mathbb{B}^{V\times V})$$
$$\Downarrow$$
$$\begin{cases} \lambda x.\mu_n^{(1)}(h,x) : X \to \mathcal{P}(\mathbb{B}^V) \\ \lambda x.\mu_n^{(2)}(h,x) : X \times \mathbb{B}^V \to \mathcal{P}(\mathbb{B}^{V\times V}) \end{cases}$$

Figure 3.29: Deconstruction in the case of directed graphs

The deconstruction of the schema μ is given in Figure 3.29, where $\mu_n = \mu_n^{(1)} \otimes \mu_n^{(2)}$ $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n))$-a.e., for all $n \in \mathbb{N}_0$. The existence of the factor $\mathbb{B}^V$ in the domain of $\mu_n^{(2)}$ is used to ensure that the above restriction on the support of each $\mu_n(h,x)$ holds. The condition on $\mu_n^{(2)}$ to ensure this is as follows.

$$\mu_n^{(2)}(h,x,\nu)(\{\varepsilon \in \mathbb{B}^{V\times V} \mid \text{for all } v,w \in V, \\ \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}) = 1, \quad (3.4.2)$$

for all $h \in H_n$, $x \in X$, $\nu \in \mathbb{B}^V$, and $n \in \mathbb{N}_0$.

Here is the proof that the Condition 3.4.2 for $\mu_n^{(2)}$ implies Condition 3.4.1 for μ_n.

$$\begin{aligned} &\mu_n(h,x)(\{(\nu,\varepsilon) \in \mathbb{B}^V \times \mathbb{B}^{V\times V} \mid \text{for all } v,w \in V, \\ &\qquad\qquad \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}) \\ &= (\mu_n^{(1)} \otimes \mu_n^{(2)})(h,x)(\{(\nu,\varepsilon) \in \mathbb{B}^V \times \mathbb{B}^{V\times V} \mid \text{for all } v,w \in V, \\ &\qquad\qquad \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}) \\ &= \int_{\mathbb{B}^V} \left(\lambda\nu. \int_{\mathbb{B}^{V\times V}} \lambda\varepsilon.\mathbf{1}_{\{(\nu,\varepsilon)\in\mathbb{B}^V\times\mathbb{B}^{V\times V} \mid \text{for all } v,w\in V,\, \varepsilon(v,w)=\mathsf{T} \text{ implies } \nu(v)=\nu(w)=\mathsf{T}\}}(\nu,\epsilon)\right. \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad d\mu_n^{(2)}(h,x,\nu)\Big)\ d\mu_n^{(1)}(h,x)\\
&= \int_{\mathbb{B}^V}\Bigg(\lambda\nu.\int_{\mathbb{B}^{V\times V}}\lambda\varepsilon.\mathbf{1}_{\mathbb{B}^V}(\nu)\ \mathbf{1}_{\{\varepsilon\in\mathbb{B}^{V\times V}\mid \text{for all } v,w\in V,\ \varepsilon(v,w)=\mathsf{T} \text{ implies } \nu(v)=\nu(w)=\mathsf{T}\}}(\epsilon)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad d\mu_n^{(2)}(h,x,\nu)\Big)\ d\mu_n^{(1)}(h,x)\\
&= \int_{\mathbb{B}^V}\Bigg(\lambda\nu.\mathbf{1}_{\mathbb{B}^V}(\nu)\ \int_{\mathbb{B}^{V\times V}}\lambda\varepsilon.\mathbf{1}_{\{\varepsilon\in\mathbb{B}^{V\times V}\mid \text{for all } v,w\in V,\ \varepsilon(v,w)=\mathsf{T} \text{ implies } \nu(v)=\nu(w)=\mathsf{T}\}}(\epsilon)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad d\mu_n^{(2)}(h,x,\nu)\Big)\ d\mu_n^{(1)}(h,x)\\
&= \int_{\mathbb{B}^V}\lambda\nu.\mathbf{1}_{\mathbb{B}^V}(\nu)\ \mu_n^{(2)}(h,x,\nu)(\{\varepsilon\in\mathbb{B}^{V\times V}\mid \text{for all } v,w\in V,\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \varepsilon(v,w)=\mathsf{T} \text{ implies } \nu(v)=\nu(w)=\mathsf{T}\})\ d\mu_n^{(1)}(h,x)\\
&= \int_{\mathbb{B}^V}\lambda\nu.\mathbf{1}_{\mathbb{B}^V}(\nu)\ d\mu_n^{(1)}(h,x) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[Condition 3.4.2]}\\
&= \mu_n^{(1)}(h,x)(\mathbb{B}^V)\\
&= 1,
\end{aligned}
$$

for all $h \in H_n$, $x \in X$, and $n \in \mathbb{N}_0$.

Each of $\mu^{(1)}$ and $\mu^{(2)}$ would likely be deconstructed further. In particular, if $\mu^{(2)}$ is deconstructed, one can ask what conditions on the factors are needed to ensure Condition 3.4.2 for $\mu_n^{(2)}$. If $V \times V$ is countably infinite, it can be written in the form $(e_m)_{m\in\mathbb{N}}$, where each e_m is a pair of vertices in V. (The case when V is finite requires some obvious modifications.) Then the deconstruction of $\mu_n^{(2)}$ is as follows:

$$
\begin{aligned}
&\mu_n^{(2,1)} : H_n \times X \times \mathbb{B}^V \to \mathcal{P}(\mathbb{B})\\
&\mu_n^{(2,2)} : H_n \times X \times \mathbb{B}^V \times \mathbb{B} \to \mathcal{P}(\mathbb{B})\\
&\qquad\qquad\vdots\\
&\mu_n^{(2,m)} : H_n \times X \times \mathbb{B}^V \times \mathbb{B}^{m-1} \to \mathcal{P}(\mathbb{B})\\
&\qquad\qquad\vdots
\end{aligned}
$$

where $\mu_n^{(2)} = \bigotimes_{m\in\mathbb{N}} \mu_n^{(2,m)}$ $\mathcal{L}((\mathbf{h}_n, \mathbf{x}_n, \boldsymbol{\nu}_n))$-a.e., by Proposition A.8.8. (Here, $\boldsymbol{\nu} : \Omega \to (\mathbb{B}^V)^{\mathbb{N}_0}$ is the relevant stochastic process.) Note that

$$\mu_n^{(2,m)} : H_n \times X \times \mathbb{B}^V \times \mathbb{B}^{m-1} \to \mathcal{P}(\mathbb{B})$$

can be written more precisely as

$$\mu_n^{(2,m)} : H_n \times X \times \mathbb{B}^V \times \mathbb{B}^{\{e_1,\ldots,e_{m-1}\}} \to \mathcal{P}(\mathbb{B}^{\{e_m\}}).$$

Suppose that, for all $m \in \mathbb{N}$, $\mu_n^{(2,m)}$ satisfies the following condition:

$$
\begin{aligned}
&\mu_n^{(2,m)}(h,x,\nu,\varepsilon)(\{b\in\mathbb{B}\mid e_m \triangleq (v,w) \text{ and}\\
&\qquad\qquad\qquad\qquad b(e_m)=\mathsf{T} \text{ implies } \nu(v)=\nu(w)=\mathsf{T}\}) = 1, \quad (3.4.3)
\end{aligned}
$$

for all $h \in H_n$, $x \in X$, $\nu \in \mathbb{B}^V$, $\varepsilon \in \mathbb{B}^{m-1}$ such that, for all $v, w \in V$, if $(v, w) \in \{e_1, \ldots, e_{m-1}\}$ and $\varepsilon(v, w) = \mathsf{T}$, then $\nu(v) = \nu(w) = \mathsf{T}$, and $n \in \mathbb{N}_0$.

Condition 3.4.3 on $\varepsilon \in \mathbb{B}^{m-1}$ means that ε is a subset of edges from $\{e_1, \ldots, e_{m-1}\}$ such that each vertex of each edge is in ν. It is now shown that, under a finiteness condition on directed graphs, Condition 3.4.3 implies Condition 3.4.2. Note that

$$\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,m)} : H_n \times X \times \mathbb{B}^V \to \mathcal{P}(\mathbb{B}^m).$$

First, it is proved by induction on m that

$$(\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,m)})(h, x, \nu)(\{\varepsilon \in \mathbb{B}^m \mid \text{for all } (v, w) \in \{e_1, \ldots, e_m\}, \\ \varepsilon(v, w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}) = 1, \quad (3.4.4)$$

for all $h \in H_n$, $x \in X$, $\nu \in \mathbb{B}^V$, and $n \in \mathbb{N}_0$. Condition 3.4.3 for $m = 1$ gives the base case. Now suppose the result holds for m. Then

$$\begin{aligned}
&(\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,m+1)})(h, x, \nu)(\{\varepsilon \in \mathbb{B}^{m+1} \mid \text{for all } (v, w) \in \{e_1, \ldots, e_{m+1}\} \\
&\qquad \varepsilon(v, w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}) \\
&= \int_{\mathbb{B}^m} \Bigg(\lambda\varepsilon. \int_{\mathbb{B}} \lambda b. \\
&\qquad \mathbf{1}_{\{\varepsilon \sqcup b \in \mathbb{B}^{m+1} \mid \text{for all } (v,w) \in \{e_1,\ldots,e_{m+1}\},\ (\varepsilon \sqcup b)(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}}(\varepsilon, b) \\
&\qquad\qquad d\mu_n^{(2,m+1)}(h, x, \nu, \varepsilon) \Bigg)\ d(\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,m)})(h, x, \nu) \\
&= \int_{\mathbb{B}^m} \Bigg(\lambda\varepsilon. \int_{\mathbb{B}} \lambda b. \\
&\qquad \mathbf{1}_{\{\varepsilon \in \mathbb{B}^m \mid \text{for all } (v,w) \in \{e_1,\ldots,e_m\},\ \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}}(\varepsilon) \\
&\qquad\qquad \mathbf{1}_{\{b \in \mathbb{B} \mid e_{m+1} \triangleq (v,w) \text{ and } b(e_m) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}}(b) \\
&\qquad\qquad\qquad d\mu_n^{(2,m+1)}(h, x, \nu, \varepsilon) \Bigg)\ d(\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,m)})(h, x, \nu) \\
&= \int_{\mathbb{B}^m} \Big(\lambda\varepsilon. \mathbf{1}_{\{\varepsilon \in \mathbb{B}^m \mid \text{for all } (v,w) \in \{e_1,\ldots,e_m\},\ \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}}(\varepsilon) \\
&\qquad \int_{\mathbb{B}} \lambda b. \mathbf{1}_{\{b \in \mathbb{B} \mid e_{m+1} \triangleq (v,w) \text{ and } b(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}}(b) \\
&\qquad\qquad d\mu_n^{(2,m+1)}(h, x, \nu, \varepsilon) \Big)\ d(\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,m)})(h, x, \nu) \\
&= \int_{\mathbb{B}^m} \lambda\varepsilon. \mathbf{1}_{\{\varepsilon \in \mathbb{B}^m \mid \text{for all } (v,w) \in \{e_1,\ldots,e_m\},\ \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}}(\varepsilon) \\
&\qquad\qquad d(\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,m)})(h, x, \nu) \\
&\qquad\qquad [\text{Condition 3.4.3 for } m+1] \\
&= 1, \qquad\qquad [\text{Induction hypothesis}]
\end{aligned}$$

for all $h \in H_n$, $x \in X$, $\nu \in \mathbb{B}^V$, and $n \in \mathbb{N}_0$. This completes the proof of Condition 3.4.4.

In practice, it is finite graphs, that is, graphs for which the set of vertices is finite, that are of most interest. For $\mu_n^{(2)}$, this means that, for all $h \in H_n$, $x \in X$, $\nu \in \mathbb{B}^V$, and $n \in \mathbb{N}_0$, there exists $M \in \mathbb{N}$ such that for $k > M$, $\mu_n^{(2,k)}$ is defined by

$$\mu_n^{(2,k)}(h, x, \nu, \varepsilon) = \delta_{\mathsf{F}},$$

for all $\varepsilon \in \mathbb{B}^{k-1}$. Note that $\mu_n^{(2,k)}$, for $k > M$, necessarily satisfies Condition 3.4.3.

Under the assumption of finite directed graphs, it is now shown that Condition 3.4.4 implies Condition 3.4.2. For this,

$$\begin{aligned}
&\mu_n^{(2)}(h,x,\nu)(\{\varepsilon \in \mathbb{B}^{V\times V} \mid \text{for all } v,w \in V, \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}) \\
&= \int_{\mathbb{B}^{V\times V}} \mathbf{1}_{\{\varepsilon\in\mathbb{B}^{V\times V} \mid \text{for all } v,w\in V,\, \varepsilon(v,w)=\mathsf{T} \text{ implies } \nu(v)=\nu(w)=\mathsf{T}\}}\, d\mu_n^{(2)}(h,x,\nu) \\
&= \int_{\mathbb{B}^M} \mathbf{1}_{\{\varepsilon\in\mathbb{B}^M \mid \text{for all } (v,w)\in\{e_1,\ldots,e_M\},\, \varepsilon(v,w)=\mathsf{T} \text{ implies } \nu(v)=\nu(w)=\mathsf{T}\}} \\
&\qquad\qquad d(\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,M)})(h,x,\nu) \\
&\qquad\qquad \text{[Proposition A.8.9]} \\
&= (\mu_n^{(2,1)} \otimes \cdots \otimes \mu_n^{(2,M)})(h,x,\nu)(\{\varepsilon \in \mathbb{B}^M \mid \text{for all } (v,w) \in \{e_1,\ldots,e_M\}, \\
&\qquad\qquad \varepsilon(v,w) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}) \\
&= 1, \qquad\qquad \text{[Condition 3.4.4]}
\end{aligned}$$

for all $h \in H_n$, $x \in X$, $\nu \in \mathbb{B}^V$, and $n \in \mathbb{N}_0$.

In summary, assuming that all graphs of interest are finite and Condition 3.4.3 holds for $\mu_n^{(2,m)}$, for all $m \in \mathbb{N}$, the schema $(\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^V \times \mathbb{B}^{V\times V}))_{n\in\mathbb{N}_0}$ has only distributions in its range that correctly model distributions over directed graphs. The $\mu_n^{(2,m)}$ are maintained by the filtering process described in Chapter 4 which will need to ensure that Condition 3.4.3 holds at all times.

A similiar analysis could be done for the set of all undirected graphs (without self-loops) whose vertices are chosen from a set V, that is,

$$\{(\nu,\varepsilon) \in \mathbb{B}^V \times \mathbb{B}^{V^{(2)}} \mid \text{for all } \{v,w\} \in V^{(2)},\, \varepsilon(\{v,w\}) = \mathsf{T} \text{ implies } \nu(v) = \nu(w) = \mathsf{T}\}.$$

A common special case is when the number of vertices is fixed. Thus suppose attention is restricted to the set of all directed graphs with a fixed set of vertices. The set of all directed graphs whose set of vertices is exactly V can be identified with $\mathbb{B}^{V\times V}$, while the corresponding schemas have the form

$$(\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^{V\times V}))_{n\in\mathbb{N}_0}.$$

Similarly, the set of all undirected graphs whose set of vertices is exactly V can be identified with $\mathbb{B}^{V^{(2)}}$, while the corresponding schemas have the form

$$(\mu_n : H_n \times X \to \mathcal{P}(\mathbb{B}^{V^{(2)}}))_{n\in\mathbb{N}_0}.$$

For undirected graphs with a fixed number of vertices, various probability distributions on graphs that were introduced by mathematicians studying random graphs and are now widely used in network science and elsewhere can be considered. (For an introduction to random graphs and citations for the random graph models mentioned here, see [118].) The first two random graph models are due to Erdős and Rényi, who essentially began the study of random graphs. Suppose the $|V| = n$. Then there are $\binom{n}{2}$ undirected graphs based on n vertices. Now suppose that each of these graphs is equally likely. This leads to the distribution having the density f given by $f(g) = \binom{n}{2}^{-1}$, for all undirected graphs g having n vertices.

3.4.5 Quotients

Quotients are useful for simplifying the codomain of a probability kernel to one where the space supporting the probability measures is simpler. However, there is a trade-off in making such simplifications that are illustrated by the following three examples. Figure 3.30 recalls how quotients work for schemas and empirical beliefs.

$$\left\{\begin{array}{l}\mu_n : H_n \times X \to \mathcal{P}(Y)\\ p : Y \to Z\end{array}\right.$$
$$\Downarrow$$
$$\mu_n/p : H_n \times X \to \mathcal{P}(Z)$$

$$\left\{\begin{array}{l}\lambda x.\mu_n(h,x) : X \to \mathcal{P}(Y)\\ p : Y \to Z\end{array}\right.$$
$$\Downarrow$$
$$\lambda x.(\mu_n/p)(h,x) : X \to \mathcal{P}(Z)$$

Figure 3.30: Quotients of a schema and an empirical belief

Example 3.4.2. The most common case of a quotient probability kernel is when the function p from Definition A.10.1 is used to define feature vectors. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ a schema. For $i = 1, \ldots, m$, let $p_i : Y \to \mathbb{B}$ be a measurable function. Now define $p : Y \to \mathbb{B}^m$ by $p = (p_1, \ldots, p_m)$. For all $y \in Y$, $(p_1(y), \ldots, p_m(y)) \in \mathbb{B}^m$ is a (Boolean) feature vector. Then

$$\mu_n/p : H_n \times X \to \mathcal{P}(\mathbb{B}^m).$$

The obvious attraction of schemas of this form is that the probability measures in the codomain are defined on the simple product space $\mathbb{B}^m$, for which most of the machinery of graphical models has been developed. However, everything depends on knowing (or being able to learn) the 'correct' p that produces the 'correct' equivalence classes on Y.

In contrast, if Y is structured, then probability measures on Y are more complicated, possibly much more complicated, than on $\mathbb{B}^n$. But, if a probability kernel $\mu : X \to \mathcal{P}(Y)$ can be learned, then it is likely to be more 'precise' than μ/p. Furthermore, dealing directly with μ avoids having to consider features at all.

The preceding example illustrates a fundamental trade-off for the approach of dealing with probability kernels for which the codomain $\mathcal{P}(Y)$ has a structured Y. The traditional approach looks for a suitable feature vector p so that the problem is reduced to dealing with a codomain of the form $\mathcal{P}(\mathbb{B}^m)$. In contrast, the approach of this book allows the possibility of either working directly with the original structured Y or finding a suitable feature vector to reduce to the $\mathbb{B}^m$ case or adopting an intermediate case that uses a p that 'partially' reduces the structure of Y. Here is an example to illustrate the latter approach.

Example 3.4.3. Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, $(Z, \mathcal{C})$, and $(W, \mathcal{D})$ be measurable spaces, and

$$(\mu_n : H_n \times X \to \mathcal{P}(\coprod_{m \in \mathbb{N}_0} Y^m \times Z \times W))_{n \in \mathbb{N}_0}$$

a schema. Thus each μ_n is a probability kernel whose codomain is probability measures on a set of triples, the first argument of which are (finite) lists whose elements are in Y, and the details of the second and third arguments are not of interest for this example. Let $\mathbb{B}^Y$ be the set of subsets of Y given the usual σ-algebra generated by the evaluation maps. Define $q : \coprod_{m \in \mathbb{N}_0} Y^m \to \mathbb{B}^Y$ by

$$q(y_1, \ldots, y_m) = \{y \mid y = y_i, \text{ for some } i \in \{1, \ldots, m\}\},$$

for all $m \in \mathbb{N}_0$ and $(y_1, \ldots, y_m) \in Y^m$. Then q is measurable, by Proposition A.1.6. Now define $p : \coprod_{m \in \mathbb{N}_0} Y^m \times Z \times W \to \mathbb{B}^Y \times Z \times W$ by $p(l, z, w) = (q(l), z, w)$, for all $l \in \coprod_{m \in \mathbb{N}_0} Y^m$, $z \in Z$, and $w \in W$. Then p is measurable and, by Proposition 3.3.5, $(\mu_n/p : H_n \times X \to \mathcal{P}(\mathbb{B}^Y \times Z \times W))_{n \in \mathbb{N}_0}$ is a schema. Each μ_n/p has a codomain that is probability measures on a set of triples the first argument of which is the set of subsets of Y.

A motivation for preferring μ_n/p to μ_n may be that distributions on the 'less structured' space $\mathbb{B}^Y \times Z \times W$ may be easier to deal with than those on $\coprod_{m \in \mathbb{N}_0} Y^m \times Z \times W$, and there may be good reason to believe that, in the context of the application, the order of elements in a list and/or their repeated occurrence may not be important.

Example 3.4.4. Example 3.4.3 shows how lists can be reduced to sets. In fact, there is an intermediate case where lists are reduced to multisets. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces and $(\mu_n : H_n \times X \to \mathcal{P}(\coprod_{m \in \mathbb{N}_0} Y^m))_{n \in \mathbb{N}_0}$ be a schema. Define $p : \coprod_{m \in \mathbb{N}_0} Y^m \to \mathbb{N}_0^Y$ by

$$p(y_1, \ldots, y_m) = \lambda y.|\{i \in \{1, \ldots, m\} \mid y = y_i\}|,$$

for all $m \in \mathbb{N}_0$ and $(y_1, \ldots, y_m) \in Y^m$. Then p is measurable, by Proposition A.1.7, and $(\mu_n/p : H_n \times X \to \mathcal{P}(\mathbb{N}_0^Y))_{n \in \mathbb{N}_0}$ is a schema, by Proposition 3.3.5.

Clearly, comparing lists, multisets, and sets, lists are the 'most' structured, recording existence, order, and multiplicity; multisets have 'intermediate' structure, recording existence and multiplicity; and sets have the 'least' structure, recording just existence. The approach of this book provides the possibility of working with probability measures on support spaces at any of these levels.

3.4.6 Function Spaces

In this case, there is a set Z and measurable spaces X and W such that schemas have the form $(\mu_n)_{n \in \mathbb{N}_0}$, where

$$\mu_n : H_n \times X \to \mathcal{P}(W^Z),$$

for all $n \in \mathbb{N}_0$. No assumption about Z is made in the following discussion, but the most interesting case is where Z is either a large finite or an infinite set; this case should be

kept in mind. Commonly in applications, X is absent at the top level, so the following discussion makes that assumption. Also, typically, W is $\mathbb{B}$, $\mathbb{N}_0$, a finite set, or $\mathbb{R}^m$, for some $m \in \mathbb{N}$.

The function space setting is illustrated in Figure 3.31. There is a stochastic process $\mathbf{k} : \Omega \to (W^Z)^{\mathbb{N}_0}$, and a probability kernel $\mu_n : H_n \to \mathcal{P}(W^Z)$ that is a regular conditional distribution of $\mathbf{k}_n$ given $\mathbf{h}_n$, for all $n \in \mathbb{N}_0$. The task is to acquire $(\mu_n)_{n \in \mathbb{N}_0}$.

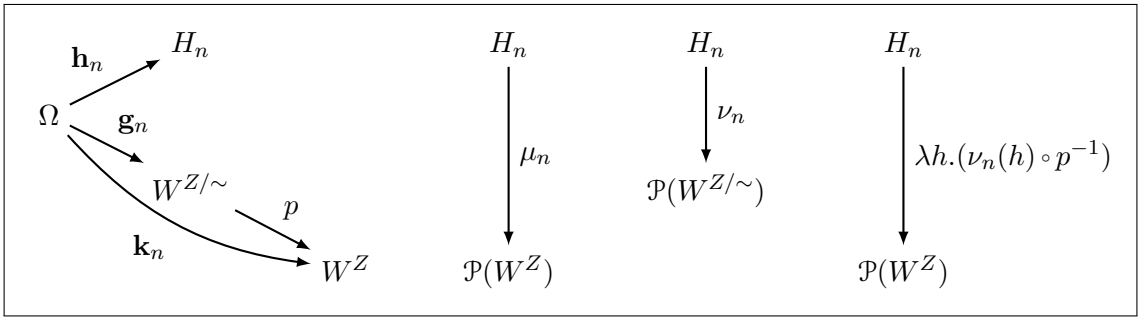

Figure 3.31: Function space setting

So consider a component $\mu_n : H_n \to \mathcal{P}(W^Z)$ of the schema. If h is the current history, this gives an empirical belief of the form $\mu_n(h) : \mathcal{P}(W^Z)$. This belief is a distribution on the set of functions from Z to W. As explained in detail below, to acquire a suitable distribution on W^Z, the key idea is to put an equivalence relation $\sim$ on Z, acquire a distribution on $W^{Z/\sim}$, and from that construct a distribution on W^Z.

If Z is infinite or very large, then working with a finite and small quotient space $Z/\sim$ is likely to be more convenient than Z. At the other extreme, it is possible for $\sim$ to be the identity relation on Z, so that $Z/\sim$ is the same as Z.

Function spaces have already been discussed for the case of sets, when $W = \mathbb{B}$, and multisets, when $W = \mathbb{N}_0$. Now two other special cases for function spaces are considered, one for classification and one for regression.

Example 3.4.5. (*Classification functions*) Consider a robot operating in some environment that includes humans. Suppose that it makes sense for the robot to model the environment as a state together with the humans as other agents. The robot may need to model the behaviour of the humans to help predict how they might react in certain circumstances. For example, if the robot moves quickly towards a particular human, the robot might want to know how the human will react. As another example, a poker-playing agent needs to model the behaviour of its opponents, for example, by modelling the circumstances under which an opponent might fold, or raise, a hand. As a third example, a virtual personal assistant may model which TV programs the user likes to watch in order to make good recommendations to the user.

All these examples involve learning the definition of a classification function. Suppose the signature of the function is $Z \to W$, where W is the set of classes. Then the setting for this problem is a schema μ, where

$$\mu_n : H_n \to \mathcal{P}(W^Z),$$

for all $n \in \mathbb{N}_0$. The associated empirical beliefs have the form $\mu_n(h_n) : \mathcal{P}(W^Z)$. Since W^Z is the set of all functions having signature $Z \to W$, an empirical belief is a distribution on

the set of functions. One can think of W^Z as the hypothesis space, the observations as the data, and the learning problem as one of finding a suitable function in W^Z. In fact, here, the problem will be to learn the *distribution* on the function space W^Z from which a particular function in W^Z can be extracted.

The function space setting is illustrated in Figure 3.31. There is a stochastic process $\mathbf{k} : \Omega \to (W^Z)^{\mathbb{N}_0}$, and a probability kernel $\mu_n : H_n \to \mathcal{P}(W^Z)$ that is a regular conditional distribution of $\mathbf{k}_n$ given $\mathbf{h}_n$, for all $n \in \mathbb{N}_0$. The task is to acquire $(\mu_n)_{n\in\mathbb{N}_0}$ (or at least some close approximation of it) and, for example, use this to classify elements of Z.

So consider a component $\mu_n : H_n \to \mathcal{P}(W^Z)$ of the schema. If h is the current history, this gives an empirical belief of the form $\mu_n(h) : \mathcal{P}(W^Z)$. This belief is a distribution on the set of functions from Z to W. Often W^Z is an hypothesis space for the problem of learning a function from Z to W from some data. Here, the distribution on the space of hypotheses will be used to find the Bayes optimal classifier [109].

To construct a suitable distribution, one can proceed as follows. Let $\sim$ be an equivalence relation on Z such that the set of equivalence classes $Z/\!\sim$ is finite (and, for practical reasons, small). Let $\pi : Z \to Z/\!\sim$ be the canonical surjection. Define $p : W^{Z/\sim} \to W^Z$ by $p(f) = f \circ \pi$, for all $f \in W^{Z/\sim}$. By Part 1 of Proposition A.10.5, p is measurable. Let $(\nu_n : H_n \to \mathcal{P}(W^{Z/\sim}))_{n\in\mathbb{N}_0}$ be the schema for $\mathbf{g}$, say. By Parts 3 and 4 of Proposition A.10.5, $(\lambda h.(\nu_n(h) \circ p^{-1}) : H_n \to \mathcal{P}(W^Z))_{n\in\mathbb{N}_0}$ is the schema for $\lambda\omega.((p \circ \mathbf{g}_n)(\omega))_{n\in\mathbb{N}_0}$. Thus, for all $n \in \mathbb{N}_0$ and $h_n \in H_n$, $\nu_n(h_n) \circ p^{-1} : \mathcal{P}(W^Z)$ is an empirical belief. The probability measure $\nu_n(h_n) \circ p^{-1}$ provides the distribution over the hypothesis space.

Let C be the set of functions in W^Z that are constant on each equivalence class in the partition of Z. Then Part 5 of Proposition A.10.5 shows that $(\nu_n(h_n) \circ p^{-1})(C) = 1$. Thus the support of the probability measure $\nu_n(h_n) \circ p^{-1}$ is the set of piecewise-constant functions in W^Z that are based on the partition induced by $\sim$. The class C provides a useful hypothesis space provided that the equivalence relation has been carefully chosen.

The next issue is how $\sim$ is chosen. The space W is now assumed to be finite to provide multiclass classification. The partition must be learned from data and, in this setting, the data comes from the history, often just the observations, up to the current time. The form that this data takes depends on the application, but it is often simply training examples of the form (z, w), where $z \in Z$ and $w \in W$. For example, for the virtual personal assistant, an action might be to recommend to the user a particular TV program z and a subsequent observation might indicate that the user watched the program to the end so that w is T. So assume that each action-observation cycle produces training examples in such a form. Now the partition induced by $\sim$ can be obtained by predicates on Z that are generated by a predicate rewrite system. As discussed in [96], to build a classifier from a set of examples, one has to find a criterion that partitions the examples into two sets which are purer in the distribution of classes that they contain than the original set; apply this process recursively on the child nodes until the leaf nodes of the tree are sufficiently pure; and then use the resulting decision tree as the induced classifier.

An important issue for decision-tree algorithms is to decide under what situations a node should be split and, if so, what predicate should be used to split it. Given that the overall aim of the algorithm is to produce a decision tree with high predictive accuracy, it is natural to propose to use accuracy as the criterion for node splitting. (The accuracy of a set of examples is the fraction of the examples in the majority class of the set; the precise definition is given in [96].) A variety of other heuristics mostly based on entropy

could also be used. The proposal here is to retain just the sequence of predicates that do the splitting, and hence define the partition, but discard the decision tree itself. The accuracy heuristic produces partitions where each equivalence class contains elements z whose corresponding w values tend to be the same. Any other heuristic with this property could be used.

Now the discussion turns to acquiring the schema $(\nu_n : H_n \to \mathcal{P}(W^{Z/\sim}))_{n\in\mathbb{N}_0}$. For this, the stochastic filtering methods of Chapter 4 can be utilized. With suitable choices of a transition model and an observation model, the schema can be filtered either by the filtering method in Section 4.1, the relevant result being Proposition 4.1.2, where the Y in that result is $W^{Z/\sim}$ here, or by the particle filtering algorithm in Section 4.3.

To extract a particular function from W^Z, one can proceed as follows. For all $z \in Z$ and $w \in W$, define $F_{z,w} : W^Z \to \mathbb{R}$ by

$$F_{z,w}(f) = \begin{cases} 1 & \text{if } f(z) = w \\ 0 & \text{otherwise,} \end{cases}$$

for all $f \in W^Z$. Note that, for all $z \in Z$, $\lambda w.F_{z,w}(f)$ is a density on W that is 1, if $f(z) = w$, and 0, otherwise. Also, for all $z \in Z$ and $w \in W$, $F_{z,w} : W^Z \to \mathbb{R}$ is measurable. Instead of f, one now effectively works with the conditional density $\lambda z.\lambda w.F_{z,w}(f) : Z \to \mathcal{D}(W)$. Then the Bayes optimal classifier $\xi : Z \to W$ is defined by

$$\xi(z) = \operatorname*{argmax}_{w\in W} \int_{W^Z} F_{z,w}\, d(\nu_n(h_n) \circ p^{-1}),$$

for all $z \in Z$. Intuitively, the Bayes optimal classifier classifies each element of Z according to the maximum of the expected vote of the hypotheses, where the vote of each hypothesis is weighted according to the distribution $\nu_n(h_n) \circ p^{-1}$ on the hypothesis space. Note that, if $z \sim z'$, then $F_{z,w}(f) = F_{z',w}(f)$, for all $w \in W$ and $f \in C$; hence $\xi(z) = \xi(z')$ and so $\xi \in C$. Thus the Bayes optimal classsifer is in the support of $\nu_n(h_n) \circ p^{-1}$.

According to Part 6 of Proposition A.10.5, $\xi : Z \to W$ is also defined by

$$\xi(z) = \operatorname*{argmax}_{w\in W} \int_{W^{Z/\sim}} F_{z,w} \circ p\, d\nu_n(h_n),$$

for all $z \in Z$. Since $W^{Z/\sim}$ is a finite product space, the integral can be computed efficiently by Monte Carlo methods and, hence, ξ can also be computed efficiently.

Finally, the discussion turns to the accuracy, or otherwise, of the distributions $\nu_n(h_n) \circ p^{-1}$, for all $n \in \mathbb{N}_0$. Given $h_n \in H_n$, the correct distribution is, of course, $\mu_n(h_n)$. So the issue is one of determining how close $\nu_n(h_n) \circ p^{-1}$ is to $\mu_n(h_n)$. Inspection of Figure 3.31 reveals that the difference between the distributions comes down to the difference between the random variables $\mathbf{k}_n$ and $p \circ \mathbf{g}_n$. If each $\mathbf{k}_n$ factors through W^F, for some finite set F, then, in principle, the equivalence relation $\sim$ can be chosen so that $p \circ \mathbf{g}_n$ is equal to $\mathbf{k}_n$ and thus, by Part 4 of Proposition A.10.5, $\nu_n(h_n) \circ p^{-1} = \mu_n(h_n)$. Otherwise, $p \circ \mathbf{g}_n$ is only an approximation to $\mathbf{k}_n$ and thus $\nu_n(h_n) \circ p^{-1}$ may only be an approximation to $\mu_n(h_n)$. Thus choosing a good partition of Z is crucial.

Further discussion of this example is given in Example 4.3.1.

Example 3.4.6. (*Regression functions*) A setting for linear regression is presented now. Let Z be $\mathbb{R}^m$, for some $m \in \mathbb{N}$, and W be $\mathbb{R}$. There is a schema $(\mu_n)_{n\in\mathbb{N}_0}$, where

$$\mu_n : H_n \to \mathcal{P}(\mathbb{R}^{\mathbb{R}^m}),$$

for all $n \in \mathbb{N}_0$, and a stochastic process $\mathbf{k} : \Omega \to (\mathbb{R}^{\mathbb{R}^m})^{\mathbb{N}_0}$. For all $n \in \mathbb{N}_0$, μ_n is a probability kernel that is a regular conditional distribution of $\mathbf{k}_n$ given $\mathbf{h}_n$. See Figure 3.32. The task is to acquire $(\mu_n)_{n\in\mathbb{N}_0}$ (or at least some close approximation of it) and, for example, use this for regression.

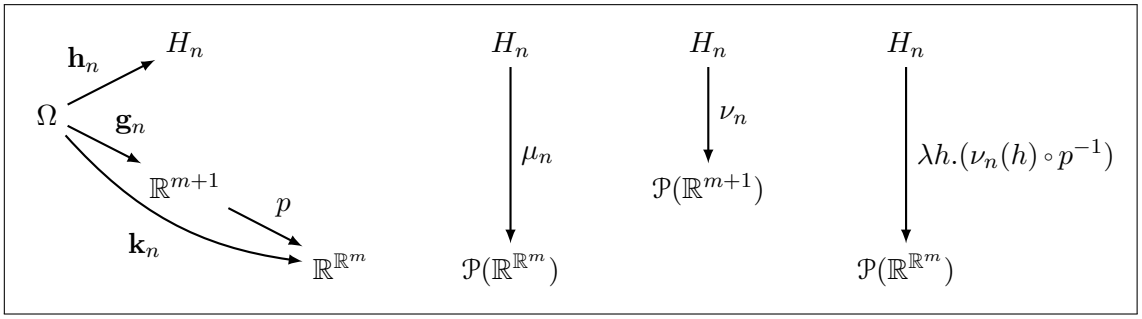

Figure 3.32: Function space setting for regression

Define $p : \mathbb{R}^{m+1} \to \mathbb{R}^{\mathbb{R}^m}$ by

$$p(a_1, \ldots, a_{m+1}) = \lambda(x_1, \ldots, x_m).(\sum_{j=1}^{m} a_j x_j + a_{m+1}),$$

for all $(a_1, \ldots, a_{m+1}) \in \mathbb{R}^{m+1}$. By Part 1 of Proposition A.10.6, p is measurable. Let $(\nu_n : H_n \to \mathcal{P}(\mathbb{R}^{m+1}))_{n\in\mathbb{N}_0}$ be the schema for $\mathbf{g}$, say. By Parts 2 and 3 of Proposition A.10.6, $(\lambda h.(\nu_n(h) \circ p^{-1}) : H_n \to \mathcal{P}(\mathbb{R}^{\mathbb{R}^m}))_{n\in\mathbb{N}_0}$ is the schema for $\lambda\omega.((p \circ \mathbf{g}_n)(\omega))_{n\in\mathbb{N}_0}$. Thus, for all $n \in \mathbb{N}_0$ and $h_n \in H_n$, $\nu_n(h_n) \circ p^{-1} : \mathcal{P}(\mathbb{R}^{\mathbb{R}^m})$ is an empirical belief. The probability measure $\nu_n(h_n) \circ p^{-1}$ provides the distribution over the hypothesis space.

Let

$$L = \{f \in \mathbb{R}^{\mathbb{R}^m} \mid f = \lambda(x_1, \ldots, x_m).(\sum_{j=1}^{m} a_j x_j + a_{m+1}), \text{ for some } (a_1, \ldots, a_{m+1}) \in \mathbb{R}^{m+1}\}.$$

Then Part 4 of Proposition A.10.6 shows that $(\nu_n(h_n) \circ p^{-1})(L) = 1$. The class L provides an hypothesis space for linear regression.

To acquire the schema $(\nu_n : H_n \to \mathcal{P}(\mathbb{R}^{m+1}))_{n\in\mathbb{N}_0}$ the filtering method of Section 4.1 is utilized. The relevant result is Proposition 4.1.2, where the Y in that result is $\mathbb{R}^{m+1}$ here.

To extract a particular function from $\mathbb{R}^{\mathbb{R}^m}$, one can proceed as follows. The Bayes optimal linear regression function $\xi : \mathbb{R}^m \to \mathbb{R}$ is defined by

$$\xi(x_1, \ldots, x_m) = \int_{\mathbb{R}^{\mathbb{R}^m}} \lambda f.f(x_1, \ldots, x_m)\, d(\nu_n(h_n) \circ p^{-1}),$$

for all $(x_1, \ldots, x_m) \in \mathbb{R}^m$. Intuitively, the value of the Bayes optimal linear regression function for each element in $\mathbb{R}^m$ is the weighted average of the values for that element for each function in the function space $\mathbb{R}^{\mathbb{R}^m}$.

According to Part 5 of Proposition A.10.6, $\xi : \mathbb{R}^m \to \mathbb{R}$ is also defined by

$$\xi(x_1, \ldots, x_m) = \int_{\mathbb{R}^{m+1}} \lambda f.f(x_1, \ldots, x_m) \circ p \; d\nu_n(h_n),$$

for all $(x_1, \ldots, x_m) \in \mathbb{R}^m$. Note how $\lambda f.f(x_1, \ldots, x_m) \circ p$ works: p takes a tuple of parameters from $\mathbb{R}^{m+1}$ and constructs the corresponding hypothesis function in $\mathbb{R}^{\mathbb{R}^m}$; then $\lambda f.f(x_1, \ldots, x_m)$ evaluates this function at $(x_1, \ldots, x_m)$. Since $\mathbb{R}^{m+1}$ is a finite product space, the integral can be computed efficiently by Monte Carlo methods and, hence, ξ can also be computed efficiently.

Finally, the discussion turns to the potential accuracy of the distributions $\nu_n(h_n) \circ p^{-1}$, for all $n \in \mathbb{N}_0$. Given $h_n \in H_n$, the correct distribution is $\mu_n(h_n)$. So the issue is one of determining how close $\nu_n(h_n) \circ p^{-1}$ is to $\mu_n(h_n)$. Figure 3.32 shows that the difference between the distributions comes down to the difference between the random variables $\mathbf{k}_n$ and $p \circ \mathbf{g}_n$. If each $\mathbf{k}_n$ factors through $\mathbb{R}^{m+1}$, then, in principle, $p \circ \mathbf{g}_n$ can equal $\mathbf{k}_n$ and thus, by Part 3 of Proposition A.10.6, $\nu_n(h_n) \circ p^{-1} = \mu_n(h_n)$. Otherwise, $p \circ \mathbf{g}_n$ is only an approximation to $\mathbf{k}_n$ and thus $\nu_n(h_n) \circ p^{-1}$ may only be an approximation to $\mu_n(h_n)$.

Bibliographical Notes

As well as being related to Piaget's concept [131] of a schema, the term 'schema' in Definition 3.1.1 is appropriate since one dictionary definition [39] of 'schema' is "an internal representation of the world; an organization of concepts and actions that can be revised by new information about the world". Thus 'schema' captures the intuition that the empirical belief $\lambda x.\mu_n(h, x)$ is contingent upon the particular history h that is observed.

Probability measures on *product* spaces are ubiquitous in applied probability, including the application of probability in artificial intelligence. On the other hand, probability measures on *sum* spaces are rarely mentioned. This is hard to understand since sum spaces are coproducts, that is, the dual notion of products, in the category of sets and hence deserve equal prominence with products, at least categorically.

Exercises

3.1 Explain the importance of the condition that schema components be regular conditional distributions in Definition 3.1.1.

3.2 Explain why agents, environments, transition models, and observation models do not satisfy Definition 3.1.1, and hence are not schemas.

3.3 Describe applications for which it would be natural for an agent to employ a schema $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$, where Y is a space of (a) sets, (b) graphs, (c) lists, (d) multisets, and (e) sequences.

3.4 Describe an application for which it would be natural for an agent to employ a schema $(\mu_n / p : H_n \times X \to \mathcal{P}(Z))_{n \in \mathbb{N}_0}$, where each component is a quotient kernel.

Chapter 4

Acquisition of Empirical Beliefs

This chapter provides a discussion of how empirical beliefs are acquired. The method is a generalization to arbitrary schemas of the filtering method of Section 2.3 for state schemas. The filter recurrence equations for filtering schemas for two settings, nonconditional and conditional, are proved. It is shown that, for every schema, the environment can be synthesized from the schema and the transition and observation models for the schema. A key result for filtering that the observation model for the nonconditional setting can be synthesized from the transition and observation models for the conditional setting (and other ingredients) is proved. Bayesian inference is shown to be a special case of filtering. Algorithms for particle filters are presented in detail. In particular, the practically important case when X is a space of parameters that are fixed but unknown is discussed in detail, as is a more general setting. Next factored particle filters that are suited to filtering in high-dimensional product spaces are presented. The development culminates with the factored conditional particle filter that can both acquire empirical beliefs and estimate parameters in high dimensions.

4.1 Nonconditional Filters

By their nature, empirical beliefs change over time. Thus, during deployment, an agent must from time to time update its empirical beliefs. This process is known as *filtering* or *tracking* an empirical belief. In this book, filtering is intended to cover the situation where an agent is trying to filter the definition of an empirical belief that is a probability kernel not just a probability measure. The discussion that follows in this section generalizes the account of filtering in Section 2.3, where it is the state distribution that is filtered. The qualification 'nonconditional' in the title of this section refers to the absence of the argument X in the domain of the schemas considered, so that the corresponding empirical beliefs are probability *measures*, and hence nonconditional in nature. Conditional filtering is covered in the next section

The setting for this section is that of schemas having the form

$$(\mu_n : H_n \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}.$$

Thus, for this case, the state space S of Section 2.3 is replaced by an arbitrary space Y. Consider an empirical belief $\mu_n(h_n) : \mathcal{P}(Y)$, where $h_n \in H_n$, that is needed to help

choose an action. As actions are performed and observations are received by the agent, the definition of $\mu_n(h_n)$ needs to be updated to a modified definition for $\mu_{n+1}(h_{n+1})$ by the filtering process. This process requires the definitions of transition and observation models which generalize those given for the case of states in Section 2.3.

To motivate these definitions, consider the state schema case. Each component of the transition model for this case has a signature of the form

$$\tau_n : A \times S \to \mathcal{P}(S).$$

This is now generalized. First, the state S is replaced by an arbitrary space Y. In addition, the simple form of observation model for states relies upon the Markov property employed in Proposition 2.3.2 which depends on particular properties of states that will not generally be true in the more general setting considered here. Without these conditional independence assumptions, the transition model becomes dependent on the history. Thus each component of the general form of a transition model has a signature of the form

$$\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y).$$

In a similar fashion, the signature of each component of the observation model becomes

$$\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O).$$

Why these are the correct forms in the more general case for the transition and observation models will become clear in Proposition 4.1.2 below. The state setting versus the more general setting of this section is summarized in Figure 4.1.

$\mu_n : H_n \to \mathcal{P}(S)$	$\mu_n : H_n \to \mathcal{P}(Y)$
$\tau_n : A \times S \to \mathcal{P}(S)$	$\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y)$
$\xi_n : S \to \mathcal{P}(O)$	$\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O)$

Figure 4.1: Comparison of transition models and observation models for state schemas versus schemas having the form $(\mu_n : H_n \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$

The values of the action and observation processes are assumed to be known to the agent, but the stochastic process on Y is not observable. The agent must estimate the values of this stochastic process using knowledge of the history, together with the transition and observation models.

Commonly, signatures for the components of the schema μ have one of the forms

$$\mu_n : H_n \to \mathcal{P}(\prod_{i\in I} Y_i),$$
$$\mu_n : H_n \to \mathcal{P}(\coprod_{i\in I} Y_i), \text{ or}$$
$$\mu_n : H_n \to \mathcal{P}(W^Z).$$

That is, the space supporting the probability measures in the codomain is either a product, a sum, or a function space, which is a special case of a product. Typically, schemas are filtered directly in the above structured form (that is, they are not first deconstructed)

and often the distributions considered for the codomains are restricted to Dirac mixture measures (that is, particles filters are employed); see Section 4.3. One situation where deconstruction is employed is when the signature has the form $\mu_n : H_n \to \mathcal{P}(X \times Y)$, where X is a parameter space and the value of the parameter is fixed but unknown; see Section 4.2.

Here are the definitions of the concepts of transition model and observation model relevant for the schemas considered in this section.

Definition 4.1.1. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(Y, \mathcal{Y})$ a measurable space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ a stochastic process. A *transition model* (*for* $\mathbf{y}$) is a sequence $\tau \triangleq (\tau_n)_{n\in\mathbb{N}}$, where

$$\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})$, for all $n \in \mathbb{N}$.

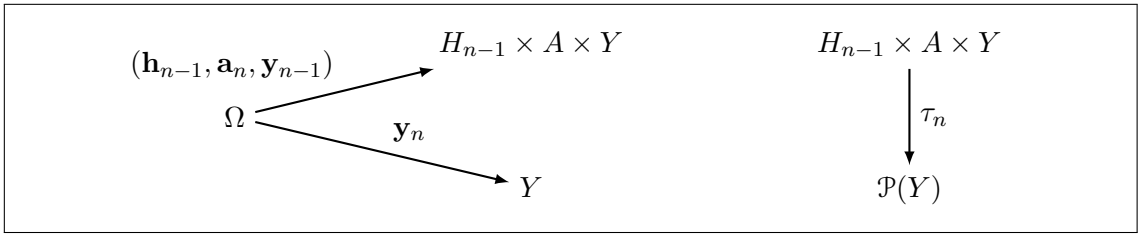

Figure 4.2: A component of a transition model

In other words, for all $n \in \mathbb{N}$, τ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})) = \lambda\omega.\tau_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. According to Proposition A.5.16, assuming that Y is a standard Borel space, for each $n \in \mathbb{N}$, such a τ_n exists and is unique $\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1}))$-a.e.

A component of a transition model for a stochastic process $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ takes as input a value in $H_{n-1} \times A \times Y$ and returns a distribution on the values in Y that could result from the transition. Here, H_{n-1} is the set of histories up to the current observation received by the agent. Having this extra argument is a requirement that comes from replacing the state space S by an arbitrary space Y; elements of Y may not force the conditional independence properties that states do.

Now comes the definition of an observation model.

Definition 4.1.2. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(Y, \mathcal{Y})$ a measurable space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ a stochastic process. An *observation model* (*for* $\mathbf{y}$) is a sequence $\xi \triangleq (\xi_n)_{n\in\mathbb{N}}$, where

$$\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O)$$

is a regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)$, for all $n \in \mathbb{N}$.

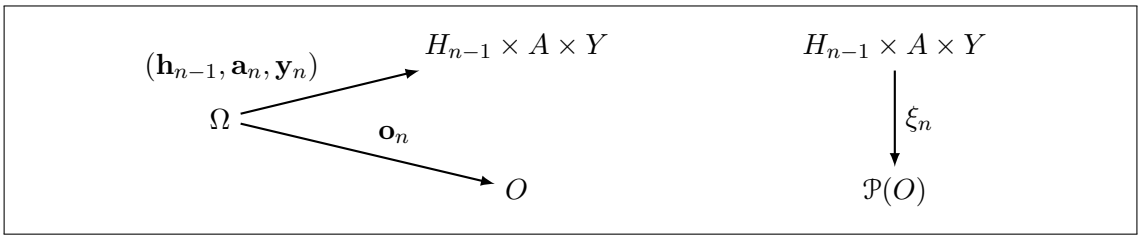

Figure 4.3: A component of an observation model

In other words, for all $n \in \mathbb{N}$, ξ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)) = \lambda\omega.\xi_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{O}$. According to Proposition A.5.16, for each $n \in \mathbb{N}$, such a ξ_n exists and is unique $\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n))$-a.e.

A component of an observation model takes as input a value in $H_{n-1} \times A \times Y$ and returns a distribution on the observations that could be received by the agent.

Some reflection will show that, in this setting, the following conditional independence property is a reasonable assumption:

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$. This property cannot be *proved* without specific assumptions about the dependency graph for the schema, indeed, for the highly complex graph which underlies all the schemas in the schema base taken together. Because the strong assumptions for state schemas are not available in this more general context, this graph has plenty of paths between typical pairs of nodes, making conditional independence properties hard to find. But it seems that for the particular property above, one can reasonably expect that $\{\mathbf{a}_{n+1}\}$ and $\{\mathbf{y}_n\}$ will be d-separated by $\{\mathbf{h}_n\}$, and hence the property will hold. This assumption can be found towards the end of the derivation of the Bayes filter in [156, Section 2.4.3]. Furthermore, in the context of a simulation for the setting of this section, the conditional independence property can be be *proved* to hold. (See the proof of Proposition 4.1.5 below.)

Here is a consequence of this conditional independence assumption.

Proposition 4.1.1. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(Y, \mathcal{Y})$ a standard Borel space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ a stochastic process, $n \in \mathbb{N}_0$, and $\mu_n : H_n \to \mathcal{P}(Y)$ a regular conditional distribution of $\mathbf{y}_n$ given $\mathbf{h}_n$. Suppose that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n).$$

Then $\lambda(h, a).\mu_n(h) : H_n \times A \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

Proof. Since μ_n is a regular conditional distribution of $\mathbf{y}_n$ given $\mathbf{h}_n$,

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid \mathbf{h}_n) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. By assumption,

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

and so, by Proposition A.6.1,

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid \mathbf{h}_n) = \mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Hence

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C) \text{ a.s.},$$

that is,

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.(\lambda(h,a).\mu_n(h))((\mathbf{h}_n, \mathbf{a}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. In other words, $\lambda(h,a).\mu_n(h) : H_n \times A \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$. □

Next comes the result giving the filter recurrence equations for schemas and empirical beliefs considered in this section.

Proposition 4.1.2. *(Filter recurrence equations for the nonconditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(Y, \mathcal{Y})$ a standard Borel space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ a stochastic process,*

$$(\mu_n : H_n \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$$

the schema for $\mathbf{y}$,

$$(\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y))_{n \in \mathbb{N}}$$

the transition model for $\mathbf{y}$,

$$(\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n \in \mathbb{N}}$$

the observation model for $\mathbf{y}$, υ_O a σ-finite measure on $\mathcal{O}$, and υ_Y a σ-finite measure on $\mathcal{Y}$. Suppose that, for all $n \in \mathbb{N}_0$, there is a conditional density $\breve{\mu}_n : H_n \to \mathcal{D}(Y)$ such that $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$ and that, for all $n \in \mathbb{N}$, there is a conditional density $\breve{\tau}_n : H_{n-1} \times A \times Y \to \mathcal{D}(Y)$ such that $\tau_n = \breve{\tau}_n \cdot \upsilon_Y$ and a conditional density $\breve{\xi}_n : H_{n-1} \times A \times Y \to \mathcal{D}(O)$ such that $\xi_n = \breve{\xi}_n \cdot \upsilon_O$. Suppose also that

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$. Then the following hold.

1. *For all $n \in \mathbb{N}_0$,*

$$\mu_{n+1} = \lambda(h,a,o).\lambda y.\breve{\xi}_{n+1}(h,a,y)(o) * \lambda(h,a,o).(\lambda(h,a).\mu_n(h) \odot \tau_{n+1})(h,a) \;\; \mathcal{L}(\mathbf{h}_{n+1})\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}(\mathbf{h}_{n+1})$*-almost all* $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\mu_{n+1}(h_{n+1}) = \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1}) * (\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)).$$

3. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}(\mathbf{h}_{n+1})$*-almost all* $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\breve{\mu}_{n+1}(h_{n+1}) = \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1}) * (\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y)) \;\; \upsilon_Y\text{-}a.e.$$

Proof. 1. Clearly, $\lambda(h,a).\breve{\mu}_n(h) : H_n \times A \to \mathcal{D}(Y)$ is a conditional density, for all $n \in \mathbb{N}_0$. Also, for all $n \in \mathbb{N}_0$,

$$\lambda(h,a).\mu_n(h) = \lambda(h,a).\breve{\mu}_n(h) \cdot \upsilon_Y.$$

To see this, since $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$, it follows that

$$\begin{aligned}
&(\lambda(h,a).\breve{\mu}_n(h) \cdot \upsilon_Y)(h,a)(C) \\
&= \int_Y \mathbf{1}_C \, \lambda(h,a).\breve{\mu}_n(h)(h,a) \, d\upsilon_Y \\
&= \int_Y \mathbf{1}_C \, \breve{\mu}_n(h) \, d\upsilon_Y \\
&= (\breve{\mu}_n \cdot \upsilon_Y)(h)(C) \\
&= \mu_n(h)(C) \\
&= \lambda(h,a).\mu_n(h)(h,a)(C),
\end{aligned}$$

for all $(h,a) \in H_n \times A$ and $C \in \mathcal{Y}$. Hence $\lambda(h,a).\mu_n(h) = \lambda(h,a).\breve{\mu}_n(h) \cdot \upsilon_Y$.

By Proposition 4.1.1, for all $n \in \mathbb{N}_0$,

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.\lambda(h,a).\mu_n(h)((\mathbf{h}_n, \mathbf{a}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Also, for all $n \in \mathbb{N}_0$, since τ_{n+1} is a regular conditional distribution,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_n)) = \lambda\omega.\tau_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Proposition A.7.15 now shows that, for all $C \in \mathcal{Y}$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1})) = \lambda\omega.(\lambda(h,a).\mu_n(h) \odot \tau_{n+1})((\mathbf{h}_n, \mathbf{a}_{n+1})(\omega))(C) \text{ a.s.}$$

Thus, for all $n \in \mathbb{N}_0$, $\lambda(h,a).\mu_n(h) \odot \tau_{n+1} : H_n \times A \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

Furthermore,

$$\begin{aligned}
&\lambda(h,a).\mu_n(h) \odot \tau_{n+1} \\
&= (\lambda(h,a).\breve{\mu}_n(h) \cdot \upsilon_Y) \odot (\breve{\tau}_{n+1} \cdot \upsilon_Y) \\
&= (\lambda(h,a).\breve{\mu}_n(h) \odot \breve{\tau}_{n+1}) \cdot \upsilon_Y. \qquad \text{[Proposition A.3.8]}
\end{aligned}$$

Hence $\lambda(h,a).\breve{\mu}_n(h) \odot \breve{\tau}_{n+1}$ is a regular conditional density of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

Next, ξ_{n+1} is an observation model; hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_{n+1})) = \lambda\omega.\xi_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_{n+1})(\omega))(B).$$

Also $\breve{\xi}_{n+1}$ is a regular conditional density.

Now consider the probability kernel

$$\lambda(h,a,o).\lambda y.\breve{\xi}_{n+1}(h,a,y)(o)*\lambda(h,a,o).(\lambda(h,a).\mu_n(h)\odot\tau_{n+1})(h,a) : H_n\times A\times O\to\mathcal{P}(Y).$$

By Proposition A.12.8, this probability kernel is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n,\mathbf{a}_{n+1},\mathbf{o}_{n+1})$. Hence

$$\lambda(h,a,o).\lambda y.\breve{\xi}_{n+1}(h,a,y)(o) * \lambda(h,a,o).(\lambda(h,a).\mu_n(h)\odot\tau_{n+1})(h,a) : H_{n+1}\to\mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $\mathbf{h}_{n+1}$. Since μ_{n+1} is by definition a regular conditional distribution of $\mathbf{y}_{n+1}$ given $\mathbf{h}_{n+1}$, it follows from the uniqueness part of Proposition A.5.16 that

$$\mu_{n+1}=\lambda(h,a,o).\lambda y.\breve{\xi}_{n+1}(h,a,y)(o)*\lambda(h,a,o).(\lambda(h,a).\mu_n(h)\odot\tau_{n+1})(h,a)\ \ \mathcal{L}(\mathbf{h}_{n+1})\text{-a.e.}$$

2. Hence, for all $n\in\mathbb{N}$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1}\triangleq(h_n,a_{n+1},o_{n+1})\in H_{n+1}$,

$$\begin{aligned}&\mu_{n+1}(h_{n+1})\\ &=\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})*(\lambda(h,a).\mu_n(h)\odot\tau_{n+1})(h_n,a_{n+1}))\\ &=\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})*(\mu_n(h_n)\odot\lambda y.\tau_{n+1}(h_n,a_{n+1},y)).\end{aligned}$$

3. For all $n\in\mathbb{N}$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1}\triangleq(h_n,a_{n+1},o_{n+1})\in H_{n+1}$,

$$\begin{aligned}&\breve{\mu}_{n+1}(h_{n+1})\cdot\upsilon_Y\\ &=\mu_{n+1}(h_{n+1})\\ &=\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})*(\mu_n(h_n)\odot\lambda y.\tau_{n+1}(h_n,a_{n+1},y))\\ &=\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ *\ ((\breve{\mu}_n(h_n)\cdot\upsilon_Y)\odot(\lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},y)\cdot\upsilon_Y))\\ &=\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ *\ ((\breve{\mu}_n(h_n)\odot\lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},y))\cdot\upsilon_Y)\\ &\qquad\qquad\text{[Proposition A.3.8]}\\ &=(\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ *\ (\breve{\mu}_n(h_n)\odot\lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},y)))\cdot\upsilon_Y.\\ &\qquad\qquad\text{[Proposition A.3.10]}\end{aligned}$$

The result now follows by Proposition A.2.11. □

Given the initial empirical belief $\mu_0(h_0)$, the recurrence equation in Part 2 of Proposition 4.1.2 enables the computation of $\mu_1(h_1)$, $\mu_2(h_2)$, and so on, where $h_0\triangleq(),h_1,h_2,\ldots$ are the successive histories.

Using the definition of the projective product, the recurrence equation for empirical beliefs (in probability measure form) is more explicitly as follows: for all $n\in\mathbb{N}_0$ and

$$\mu_{n+1} = \underbrace{\lambda(h,a,o).\lambda y.\breve{\xi}_{n+1}(h,a,y)(o)}_{\textit{observation update}} * \underbrace{\lambda(h,a,o).(\lambda(h,a).\mu_n(h) \odot \tau_{n+1})(h,a)}_{\textit{transition update}}$$

$$\mu_{n+1}(h_{n+1}) = \underbrace{\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})}_{\textit{observation update}} * \underbrace{(\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},y))}_{\textit{transition update}}$$

$$\breve{\mu}_{n+1}(h_{n+1}) = \underbrace{\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})}_{\textit{observation update}} * \underbrace{(\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},y))}_{\textit{transition update}}$$

Figure 4.4: Recurrence equations for the nonconditional case

$\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\mu_{n+1}(h_{n+1}) \\
=\ &\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1}) \ * \ (\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},y)) \\
=\ &\lambda B.\frac{\int_Y \mathbf{1}_B \lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ d(\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},y))}{\int_Y \lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ d(\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},y))} \\
=\ &\lambda B.\frac{\int_Y \left(\lambda y'.\int_Y \mathbf{1}_B \lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ d\tau_{n+1}(h_n,a_{n+1},y')\right)\ d\mu_n(h_n)}{\int_Y \left(\lambda y'.\int_Y \lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ d\tau_{n+1}(h_n,a_{n+1},y')\right)\ d\mu_n(h_n)}.
\end{aligned}$$

[Proposition A.7.8]

Thus, for all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\mu_{n+1}(h_{n+1}) = \lambda B.\frac{\int_Y \left(\lambda y'.\int_Y \mathbf{1}_B \lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ d\tau_{n+1}(h_n,a_{n+1},y')\right)\ d\mu_n(h_n)}{\int_Y \left(\lambda y'.\int_Y \lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ d\tau_{n+1}(h_n,a_{n+1},y')\right)\ d\mu_n(h_n)}.$$

Similarly, the recurrence equation for empirical beliefs in density form is more explicitly as follows: for all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$, υ_S-almost everywhere,

$$\begin{aligned}
&\breve{\mu}_{n+1}(h_{n+1}) \\
=\ &\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1}) \ * \ (\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},y)) \\
=\ &\frac{\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ (\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},y))}{\int_Y \lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},y)(o_{n+1})\ (\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},y))\ d\upsilon_Y}
\end{aligned}$$

$$= \frac{\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ d\upsilon_Y}{\int_Y(\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ d\upsilon_Y)\ d\upsilon_Y}$$

Thus, for all $n \in \mathbb{N}_0$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\breve{\mu}_{n+1}(h_{n+1}) = \frac{\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ d\upsilon_Y}{\int_Y(\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ d\upsilon_Y)\ d\upsilon_Y}$$

υ_S-a.e.

The two preceding recurrence equations in explicit form are illustrated in Figure 4.5. It is worth comparing the equations in Figure 4.5 to the corresponding equations for state distributions in Figure 2.9.

$$\mu_{n+1}(h_{n+1}) = \lambda B.\frac{\int_Y \left(\lambda y'.\int_Y \mathbf{1}_B \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ d\tau_{n+1}(h_n, a_{n+1}, y')\right)\ d\mu_n(h_n)}{\int_Y \left(\lambda y'.\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ d\tau_{n+1}(h_n, a_{n+1}, y')\right)\ d\mu_n(h_n)}$$

$$\breve{\mu}_{n+1}(h_{n+1}) = \frac{\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ d\upsilon_Y}{\int_Y(\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ d\upsilon_Y)\ d\upsilon_Y}$$

Figure 4.5: Recurrence equations for the empirical beliefs of this section in explicit form

Example 4.1.1. Under the conditions of Proposition 4.1.2, suppose that $Y = \prod_{i=1}^m Y_i$ and $O = \prod_{i=1}^m O_i$, where $(Y_i, \mathcal{Y}_i)$ is a standard Borel space and $(O_i, \mathcal{O}_i)$ an observation space, for $i = 1, \ldots, m$. For $i = 1, \ldots, m$, define $\mathbf{y}^{(i)} : \Omega \to Y_i^{\mathbb{N}_0}$ by $\mathbf{y}^{(i)} = (\mathbf{y}_n^{(i)})_{n\in\mathbb{N}_0}$, where $\mathbf{y}_n^{(i)} : \Omega \to Y_i$ is defined by $\mathbf{y}_n^{(i)} = \lambda(y_1, \ldots, y_m).y_i \circ \mathbf{y}_n$, for all $n \in \mathbb{N}_0$. Clearly, each $\mathbf{y}^{(i)} : \Omega \to Y_i^{\mathbb{N}_0}$ is a stochastic process. This example is concerned with the conditions under which an empirical belief for $\mathbf{y}$ factorizes into a product of empirical beliefs for $\mathbf{y}^{(i)}$, for $i = 1, \ldots, m$. Sufficient conditions for doing this are obvious and strong, but sometimes can be realized, especially if suitable approximations of the transition and/or observation models are justified.

The ingredients of Proposition 4.1.2 need to be set up for each Y_i and O_i. Suppose that $\upsilon_Y = \bigotimes_{i=1}^m \upsilon_{Y_i}$, where υ_{Y_i} is a σ-finite measure on Y_i, for $i = 1, \ldots, m$. For $i = 1, \ldots, m$, there is an observation process $\mathbf{o}^{(i)} : \Omega \to O_i^{\mathbb{N}}$ defined using the observation process $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ in the obvious way. For $i = 1, \ldots, m$, let $H_n^{(i)} \triangleq A \times O_i \times \cdots \times A \times O_i$, where there are n occurrences of A and n occurrences of O_i. Then $H_n^{(i)}$ is the set of histories of

action-observation cycles up until the end of the nth cycle relevant for the ith factor. If $i \in \{1, \ldots, m\}$ and $h_n \in H_n$, where

$$h_n = (a_1, (o_{1,1}, \ldots, o_{1,m}), a_2, (o_{2,1}, \ldots, o_{2,m}), \ldots, a_n, (o_{n,1}, \ldots, o_{n,m})),$$

then

$$h_n^{(i)} \triangleq (a_1, o_{1,i}, a_2, o_{2,i}, \ldots, a_n, o_{n,i}).$$

Note that $h_n^{(i)} \in H_n^{(i)}$. The action space $(A, \mathcal{A})$ is shared for each factor.

First consider the factorization of (the density form of) the transition model $(\breve{\tau}_n : H_{n-1} \times A \times Y \to \mathcal{D}(Y))_{n\in\mathbb{N}}$ for $\mathbf{y}$. Suppose that, for all $n \in \mathbb{N}$,

$$\breve{\tau}_n(h_{n-1}, a_n, (y'_1, \ldots, y'_m)) = \lambda(y_1, \ldots, y_m). \prod_{i=1}^{m} \breve{\tau}_n^{(i)}(h_{n-1}^{(i)}, a_n, y'_i)(y_i),$$

for all $h_{n-1} \in H_{n-1}$, $a_n \in A$, and $(y'_1, \ldots, y'_m) \in Y$, and where $(\breve{\tau}_n^{(i)} : H_{n-1}^{(i)} \times A \times Y_i \to \mathcal{D}(Y_i))_{n\in\mathbb{N}}$ is (the density form of) the transition model for $\mathbf{y}^{(i)}$, for $i = 1, \ldots, m$.

Next consider the factorization of (the density form of) the observation model $(\breve{\xi}_n : H_{n-1} \times A \times Y \to \mathcal{D}(O))_{n\in\mathbb{N}}$ for $\mathbf{y}$. Suppose that, for all $n \in \mathbb{N}$,

$$\breve{\xi}_n(h_{n-1}, a_n, (y_1, \ldots, y_m)) = \lambda(o_1, \ldots, o_m). \prod_{i=1}^{m} \breve{\xi}_n^{(i)}(h_{n-1}^{(i)}, a_n, y_i)(o_i),$$

for all $h_{n-1} \in H_{n-1}$, $a_n \in A$, and $(y_1, \ldots, y_m) \in Y$, and where $(\breve{\xi}_n^{(i)} : H_{n-1}^{(i)} \times A \times Y_i \to \mathcal{D}(O_i))_{n\in\mathbb{N}}$ is (the density form of) the observation model for $\mathbf{y}^{(i)}$, for $i = 1, \ldots, m$.

Now the factorization of empirical beliefs obtained from (the density form of) the schema $(\breve{\mu}_n : H_n \to \mathcal{D}(Y))_{n\in\mathbb{N}}$ for $\mathbf{y}$ can be proved. To get started, suppose that the empirical belief factorizes at time 0, so that

$$\breve{\mu}_0(()) = \lambda(y_1, \ldots, y_m). \prod_{i=1}^{m} \breve{\mu}_0^{(i)}(())(y_i),$$

where $\breve{\mu}_0^{(i)}(())$ is the empirical belief at time 0 for $\mathbf{y}^{(i)}$, for $i = 1, \ldots, m$. It will be proved by induction that

$$\breve{\mu}_n(h_n) = \lambda(y_1, \ldots, y_m). \prod_{i=1}^{m} \breve{\mu}_n^{(i)}(h_n^{(i)})(y_i),$$

for all $h_n \in H_n$, and where $(\breve{\mu}_n^{(i)} : H_n^{(i)} \to \mathcal{D}(Y_i))_{n\in\mathbb{N}}$ is (the density form of) the schema for $\mathbf{y}^{(i)}$. The base case is obvious.

Suppose now the result holds for n. The numerator of the right hand side of the second

equation in Figure 4.5 is evaluated, as follows.

$$
\begin{aligned}
&\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o)\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ dv_Y \\
&= \lambda(y_1,\ldots,y_m).\breve{\xi}_{n+1}(h_n, a_{n+1}, (y_1,\ldots,y_m))((o_1,\ldots,o_m)) \\
&\quad \lambda(y_1,\ldots,y_m).\int_Y \lambda(y'_1,\ldots,y'_m).\breve{\tau}_{n+1}(h_n, a_{n+1}, (y'_1,\ldots,y'_m))((y_1,\ldots,y_m))\ \breve{\mu}_n(h_n)\ dv_Y \\
&= \lambda(y_1,\ldots,y_m).\prod_{i=1}^m \breve{\xi}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y_i)(o_i) \\
&\quad \lambda(y_1,\ldots,y_m).\int_Y \lambda(y'_1,\ldots,y'_m).\left[\prod_{i=1}^m \breve{\tau}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y'_i)(y_i)\ \prod_{i=1}^m \breve{\mu}^{(i)}_n(h_n^{(i)})(y'_i)\right] dv_Y \\
&\qquad \text{[Definitions of the transition and observation models, and the induction hypothesis]} \\
&= \lambda(y_1,\ldots,y_m).\prod_{i=1}^m \breve{\xi}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y_i)(o_i) \\
&\quad \lambda(y_1,\ldots,y_m).\int_Y \lambda(y'_1,\ldots,y'_m).\left[\prod_{i=1}^m \breve{\tau}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y'_i)(y_i)\ \breve{\mu}^{(i)}_n(h_n^{(i)})(y'_i)\right] dv_Y \\
&= \lambda(y_1,\ldots,y_m).\left(\prod_{i=1}^m \breve{\xi}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y_i)(o_i)\right. \\
&\qquad \left.\int_{\prod_{i=1}^m Y_i} \lambda(y'_1,\ldots,y'_m).\left[\prod_{i=1}^m \breve{\tau}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y'_i)(y_i)\ \breve{\mu}^{(i)}_n(h_n^{(i)})(y'_i)\right] d\bigotimes_{i=1}^m v_{Y_i}\right) \\
&= \lambda(y_1,\ldots,y_m).\left(\prod_{i=1}^m \breve{\xi}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y_i)(o_i)\right. \\
&\quad \left.\int_{Y_m}\left(\lambda y'_n.\ \ldots\int_{Y_1}\lambda y'_1.\left[\prod_{i=1}^m \breve{\tau}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y'_i)(y_i)\ \breve{\mu}^{(i)}_n(h_n^{(i)})(y'_i)\right] dv_{Y_1}\ \ldots\right)\ dv_{Y_m}\right) \\
&\qquad\qquad \text{[Fubini theorem (Proposition A.2.16)]} \\
&= \lambda(y_1,\ldots,y_m).\left(\prod_{i=1}^m \breve{\xi}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y_i)(o_i)\right. \\
&\qquad\qquad \left.\prod_{i=1}^m\left(\int_{Y_i}\lambda y'_i.\breve{\tau}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y'_i)(y_i)\ \breve{\mu}^{(i)}_n(h_n^{(i)})(y'_i)\ dv_{Y_i}\right)\right) \\
&\qquad\qquad \text{[Factored form of integrand]} \\
&= \lambda(y_1,\ldots,y_m). \\
&\quad \prod_{i=1}^m\left(\breve{\xi}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y_i)(o_i)\ \int_{Y_i}\lambda y'_i.\breve{\tau}^{(i)}_{n+1}(h_n^{(i)}, a_{n+1}, y'_i)(y_i)\ \breve{\mu}^{(i)}_n(h_n^{(i)})(y'_i)\ dv_{Y_i}\right).
\end{aligned}
$$

Hence

$$\breve{\mu}_{n+1}(h_{n+1})$$

$$= \frac{\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ d\upsilon_Y}{\int_Y (\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\ \breve{\mu}_n(h_n)\ d\upsilon_Y)\ d\upsilon_Y}$$

$$= \frac{\lambda(y_1,\ldots,y_m).\prod_{i=1}^m \left(\breve{\xi}^{(i)}_{n+1}(h^{(i)}_n, a_{n+1}, y_i)(o_i)\ \int_{Y_i} \lambda y_i'.\breve{\tau}^{(i)}_{n+1}(h^{(i)}_n, a_{n+1}, y_i')(y_i)\ \breve{\mu}^{(i)}_n(h^{(i)}_n)(y_i')\ d\upsilon_{Y_i}\right)}{\int_Y \lambda(y_1,\ldots,y_m).\prod_{i=1}^m \left(\breve{\xi}^{(i)}_{n+1}(h^{(i)}_n, a_{n+1}, y_i)(o_i)\ \int_{Y_i} \lambda y_i'.\breve{\tau}^{(i)}_{n+1}(h^{(i)}_n, a_{n+1}, y_i')(y_i)\ \breve{\mu}^{(i)}_n(h^{(i)}_n)(y_i')\ d\upsilon_{Y_i}\right)\ d\upsilon_Y}$$

$$= \frac{\lambda(y_1,\ldots,y_m).\prod_{i=1}^m \left(\breve{\xi}^{(i)}_{n+1}(h^{(i)}_n, a_{n+1}, y_i)(o_i)\ \int_{Y_i} \lambda y_i'.\breve{\tau}^{(i)}_{n+1}(h^{(i)}_n, a_{n+1}, y_i')(y_i)\ \breve{\mu}^{(i)}_n(h^{(i)}_n)(y_i')\ d\upsilon_{Y_i}\right)}{\prod_{i=1}^m \int_{Y_i} \left(\lambda y_i.\breve{\xi}^{(i)}_{n+1}(h^{(i)}_n, a_{n+1}, y_i)(o_i)\ \lambda y_i.\int_{Y_i} \lambda y_i'.\breve{\tau}^{(i)}_{n+1}(h^{(i)}_n, a_{n+1}, y_i')(y_i)\ \breve{\mu}^{(i)}_n(h^{(i)}_n)(y_i')\ d\upsilon_{Y_i}\right)\ d\upsilon_{Y_i}}$$

[Fubini theorem (Proposition A.2.16)]

$$= \lambda(y_1,\ldots,y_m).\prod_{i=1}^m \breve{\mu}^{(i)}_{n+1}(h^{(i)}_{n+1})(y_i). \qquad \text{[Figure 4.5]}$$

This completes the induction argument.

Now the two common cases of sums and products in the codomain of schemas are examined in turn.

Example 4.1.2. Consider a schema μ having the form

$$(\mu_n : H_n \to \mathcal{P}(\coprod_{i\in I} Y_i))_{n\in\mathbb{N}_0},$$

This is a suitable form, for example, for (top-level) schemas whose codomains are probability measures over lists, as in Section 3.4.3. Consider the problem of acquiring such schemas. One approach is to directly filter the schema using the results of this section. For this, the transition model has the form

$$(\tau_n : H_{n-1} \times A \times \coprod_{i\in I} Y_i \to \mathcal{P}(\coprod_{i\in I} Y_i))_{n\in\mathbb{N}}$$

and the observation model has the form

$$(\xi_n : H_{n-1} \times A \times \coprod_{i\in I} Y_i \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

Alternatively, one can deconstruct the schema and then filter. In this case, it is convenient to assume the schema μ has the form

$$(\breve{\chi}_n \bullet \bigoplus_{i\in I} \mu^{(i)}_n : H_n \to \mathcal{P}(\coprod_{i\in I} Y_i))_{n\in\mathbb{N}_0},$$

for some $\chi \triangleq (\chi_n : H_n \to \mathcal{P}(I))_{n\in\mathbb{N}_0}$ and $\mu^{(i)} \triangleq (\mu^{(i)}_n : H_n \to \mathcal{P}(Y_i))_{n\in\mathbb{N}_0}$, for all $i \in I$ (as in Proposition 3.2.3). The task now is to filter the weight function and each of the addenda.

To filter χ, one needs a transition model

$$(\eta_n : H_{n-1} \times A \times I \to \mathcal{P}(I))_{n\in\mathbb{N}},$$

and an observation model

$$(\zeta_n : H_{n-1} \times A \times I \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

Then Proposition 4.1.2 shows that, for all $n \in \mathbb{N}_0$,

$$\chi_{n+1} = \lambda(h,a,o).\lambda i.\breve{\xi}_{n+1}(h,a,i)(o) * \lambda(h,a,o).(\chi_n^\dagger \odot \eta_{n+1})(h,a) \;\; \mathcal{L}(\mathbf{h}_{n+1})\text{-a.e.}$$

To filter $\mu^{(i)}$, for all $i \in I$, one needs a transition model

$$(\tau_n^{(i)} : H_{n-1} \times A \times Y_i \to \mathcal{P}(Y_i))_{n\in\mathbb{N}}$$

and an observation model

$$(\xi_n^{(i)} : H_{n-1} \times A \times Y_i \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

Then Proposition 4.1.2 shows that, for all $i \in I$ and $n \in \mathbb{N}_0$,

$$\mu_{n+1}^{(i)} = \lambda(h,a,o).\lambda y_i.\breve{\xi}_{n+1}^{(i)}(h,a,y_i)(o) * \lambda(h,a,o).(\mu_n^{(i)\dagger} \odot \tau_{n+1}^{(i)})(h,a) \;\; \mathcal{L}(\mathbf{h}_{n+1})\text{-a.e.}$$

Now, by Propositions 3.2.3 and A.5.16,

$$\mu_{n+1} = \breve{\chi}_{n+1} \bullet \bigoplus_{i\in I} \mu_{n+1}^{(i)} \;\; \mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1})\text{-a.e.},$$

for all $n \in \mathbb{N}_0$. Hence the task of filtering each μ_n is thus reduced to filtering χ_n and each $\mu_n^{(i)}$ and then forming the weighted sum of the results of the filtering. Also recall from the discussion in Section 3.4.3 that, in practice, it is only ever necessary to deal with finitely many of the $\mu^{(i)}$.

There is a similar development to the above for the corresponding empirical beliefs.

Example 4.1.3. In contrast to Example 4.1.2, the product case is now considered. Let the schema μ have the form

$$(\mu_n : H_n \to \mathcal{P}(\prod_{i\in I} Y_i))_{n\in\mathbb{N}_0}.$$

One approach is to directly filter the schema using the results of this section. For this, the transition model has the form

$$(\tau_n : H_{n-1} \times A \times \prod_{i\in I} Y_i \to \mathcal{P}(\prod_{i\in I} Y_i))_{n\in\mathbb{N}}$$

and the observation model has the form

$$(\xi_n : H_{n-1} \times A \times \prod_{i\in I} Y_i \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

As an alternative, one can first deconstruct μ and then filter. Suppose that either $I = \{1,\ldots,m\}$, for some $m \in \mathbb{N}$, or $I = \mathbb{N}$. Then deconstruction gives schemas

$$(\mu_n^{(i)} : H_n \times \prod_{j=1}^{i-1} Y_j \to \mathcal{P}(Y_i))_{n\in\mathbb{N}_0}$$

such that

$$\mu_n = \bigotimes_{i \in I} \mu_n^{(i)} \;\; \mathcal{L}(\mathbf{h}_n)\text{-a.e.},$$

for all $n \in \mathbb{N}_0$. Consider now the task of filtering the $\mu^{(i)}$. Differently to Example 4.1.2, there is an additional factor $\prod_{j=1}^{i-1} Y_j$ introduced into the domains of the $\mu^{(i)}$ by the deconstruction. This means that, for $i > 1$, the situation is no longer the nonconditional case of this section but, instead, the conditional case of the next section. But this creates an impediment: the proposition giving the filter recurrence equations for the conditional case (Proposition 4.2.3) has a condition on this extra domain argument that, while useful for some important cases like fixed but unknown parameters, is not generally satisfied.

The upshot is that, in the product case, deconconstruction followed by filtering is only useful for certain special situations such as learning fixed parameters. More generally useful is either filtering without deconstructing as noted above or else employing a particle filter instead, as in Section 4.3, that does not need deconstruction and can be applied directly in structured cases.

From the ingredients of Proposition 4.1.2, the environment can be synthesized.

Proposition 4.1.3. *(Environment synthesis for the nonconditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(Y, \mathcal{Y})$ a standard Borel space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ a stochastic process,*

$$(\mu_n : H_n \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$$

the schema for $\mathbf{y}$,

$$(\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y))_{n \in \mathbb{N}}$$

the transition model for $\mathbf{y}$,

$$(\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n \in \mathbb{N}}$$

the observation model for $\mathbf{y}$, υ_O a σ-finite measure on $\mathcal{O}$, and υ_Y a σ-finite measure on $\mathcal{Y}$. Let $(\Xi_n : H_{n-1} \times A \to \mathcal{P}(O))_{n \in \mathbb{N}}$ be the environment for $\mathbf{a}$ and $\mathbf{o}$. Suppose that, for all $n \in \mathbb{N}_0$, there is a conditional density $\breve{\mu}_n : H_n \to \mathcal{D}(Y)$ such that $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$. Suppose that, for all $n \in \mathbb{N}$, there is a conditional density $\breve{\Xi}_n : H_{n-1} \times A \to \mathcal{D}(O)$ such that $\Xi_n = \breve{\Xi}_n \cdot \upsilon_O$, a conditional density $\breve{\tau}_n : H_{n-1} \times A \times Y \to \mathcal{D}(Y)$ such that $\tau_n = \breve{\tau}_n \cdot \upsilon_Y$, and a conditional density $\breve{\xi}_n : H_{n-1} \times A \times Y \to \mathcal{D}(O)$ such that $\xi_n = \breve{\xi}_n \cdot \upsilon_O$. Suppose also that

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$. Then the following hold.

1. *For all $n \in \mathbb{N}_0$,*

$$\Xi_{n+1} = (\lambda(h, a).\mu_n(h) \odot \tau_{n+1}) \odot \xi_{n+1} \;\; \mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$,

$$\Xi_{n+1}(h_n, a_{n+1}) = (\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, y).$$

3. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$,

$$\breve{\Xi}_{n+1}(h_n, a_{n+1}) = (\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y) \ \upsilon_O\text{-a.e.}$$

4. *For all* $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$, *and all* $C \in \mathcal{O}$,

$$\Xi_{n+1}(h_n, a_{n+1})(C) = \int_Y \lambda y.\xi_{n+1}(h_n, a_{n+1}, y)(C)\ d(\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)).$$

5. *For all* $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$, *and* υ_O*-almost all* $o \in O$,

$$\breve{\Xi}_{n+1}(h_n, a_{n+1})(o) = \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o)\ d(\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)).$$

Proof. 1. By Proposition 4.1.1, $\lambda(h, a).\mu_n(h) : H_n \times A \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. Then, by Proposition A.7.15, for all $n \in \mathbb{N}_0$, $\lambda(h, a).\mu_n(h) \odot \tau_{n+1} : H_n \times A \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$. By Proposition A.7.15 again, $(\lambda(h, a).\mu_n(h) \odot \tau_{n+1}) \odot \xi_{n+1} : H_n \times A \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. However, from Definition 2.2.2, $\Xi_{n+1} : H_n \times A \to \mathcal{P}(O)$ is also a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. The result now follows from the uniqueness part of Proposition A.5.16.

2. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$,

$$\begin{aligned}
&\Xi_{n+1}(h_n, a_{n+1}) \\
&= ((\lambda(h, a).\mu_n(h) \odot \tau_{n+1}) \odot \xi_{n+1})(h_n, a_{n+1}) \\
&= (\lambda(h, a).\mu_n(h) \odot \tau_{n+1})(h_n, a_{n+1}) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, y) \\
&= (\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, y).
\end{aligned}$$

3. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$,

$$\begin{aligned}
&\breve{\Xi}_{n+1}(h_n, a_{n+1}) \cdot \upsilon_O \\
&= \Xi_{n+1}(h_n, a_{n+1}) \\
&= (\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, y) \\
&= ((\breve{\mu}_n(h_n) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y) \cdot \upsilon_Y)) \odot (\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y) \cdot \upsilon_O) \\
&= ((\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y) \cdot \upsilon_O) \\
&\qquad\qquad \text{[Proposition A.3.8]} \\
&= ((\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)) \cdot \upsilon_O. \\
&\qquad\qquad \text{[Proposition A.3.8]}
\end{aligned}$$

The result now follows by Proposition A.2.11.

4. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$, and all $C \in \mathcal{O}$,

$$\begin{aligned}
&\Xi_{n+1}(h_n, a_{n+1})(C) \\
&= ((\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, y))(C) \\
&= \int_Y \lambda y.\xi_{n+1}(h_n, a_{n+1}, y)(C)\ d(\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)).
\end{aligned}$$

5. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$, and υ_O-almost all $o \in O$,

$$\begin{aligned}
&\breve{\Xi}_{n+1}(h_n, a_{n+1})(o) \\
&= ((\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y))(o) \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o)\ (\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y))\ d\upsilon_Y \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o)\ d((\breve{\mu}_n(h_n) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y)) \cdot \upsilon_Y) \\
&\qquad\qquad \text{[Proposition A.3.3]} \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o)\ d((\breve{\mu}_n(h_n) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, y) \cdot \upsilon_Y)) \\
&\qquad\qquad \text{[Proposition A.3.8]} \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o)\ d(\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)).
\end{aligned}$$

□

Part 1 of Proposition 4.1.3 shows that, for every schema $(\mu_n : H_n \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$, the environment can be synthesized from the schema and the related transition and observation models. Part 2 shows that, given the current history h_n and the action a_{n+1} just made, the agent can compute the probability measure giving the distribution for the next observation o_{n+1} from the empirical belief $\mu_n(h_n)$ and the transition and observation models. Part 3 is similar, but for densities. Parts 4 and 5 give convenient ways of calculating $\Xi_{n+1}(h_n, a_{n+1})(C)$, for $C \in \mathcal{O}$, and $\breve{\Xi}_{n+1}(h_n, a_{n+1})(o)$, for $o \in O$.

The environment is the observation model for a particular (trivial) schema.

Example 4.1.4. Proposition 4.1.3 shows that the environment can be synthesized from any schema and the related transition and observation models. Conversely, from the environment, one can construct a particular schema and the related transition and observation models. Let $(\Xi_n : H_{n-1} \times A \to \mathcal{P}(O))_{n \in \mathbb{N}}$ be the environment and Y a singleton set $\{y_0\}$. For all $n \in \mathbb{N}_0$, let $\mu_n : H_n \to \mathcal{P}(Y)$ be the only possible function that takes any history to the unique probability measure on Y. Then $(\mu_n : H_n \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$ is the schema for $\mathbf{y}$, where $\mathbf{y}$ is the only possible function from Ω to $Y^{\mathbb{N}_0}$.

For all $n \in \mathbb{N}$, define $\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y)$ to be the only possible function that takes any history, action, and y_0 to the unique probability measure on Y. Then $(\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y))_{n \in \mathbb{N}}$ is the transition model for $\mathbf{y}$. For all $n \in \mathbb{N}$, define $\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O)$ by $\xi_n(h, a, y_0) = \Xi_n(h, a)$, for all $h \in H_n$ and $a \in A$. Then $(\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n \in \mathbb{N}}$ is the observation model for $\mathbf{y}$. By Proposition 4.1.3,

$\Xi_{n+1} = (\lambda(h,a).\mu_n(h) \odot \tau_{n+1}) \odot \xi_{n+1}$ $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-a.e.. In effect, Ξ is the observation model for $\mathbf{y}$.

By analogy with the case of states, suppose now that the conditional independence assumption

$$\sigma(\mathbf{o}_n) \perp\!\!\!\perp_{\sigma(\mathbf{y}_n)} \sigma(\mathbf{h}_{n-1}, \mathbf{a}_n),$$

for all $n \in \mathbb{N}$, is satisfied. For all $n \in \mathbb{N}$, let $\xi'_n : Y \to \mathcal{P}(O)$ be the regular conditional distribution of $\mathbf{o}_n$ given $\mathbf{y}_n$. Hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid \mathbf{y}_{n+1}) = \lambda\omega.\xi'_{n+1}(\mathbf{y}_{n+1}(\omega))(B) \text{ a.s.}$$

Consider $\lambda(h,a,y).\xi'_{n+1}(y) : H_n \times A \times Y \to \mathcal{P}(O)$. Now, for all $B \in \mathcal{O}$,

$$\lambda\omega.\xi'_{n+1}(\mathbf{y}_{n+1}(\omega))(B) = \lambda\omega.(\lambda(h,a,y).\xi'_{n+1}(y))((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_{n+1})(\omega))(B).$$

Since $\sigma(\mathbf{o}_{n+1}) \perp\!\!\!\perp_{\sigma(\mathbf{y}_{n+1})} \sigma(\mathbf{h}_n, \mathbf{a}_{n+1})$, it follows by Proposition A.6.1 that

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid \mathbf{y}_{n+1}) = \mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_{n+1})) \text{ a.s.},$$

for all $B \in \mathcal{O}$. Hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_{n+1})) = \lambda\omega.(\lambda(h,a,y).\xi'_{n+1}(y))((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_{n+1})(\omega))(B).$$

This means that $\lambda(h,a,y).\xi'_{n+1}(y)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{a}_{n+1}, \mathbf{h}_n, \mathbf{y}_{n+1})$. It follows that

$$\xi_{n+1} = \lambda(h,a,y).\xi'_{n+1}(y) \;\; \mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_{n+1}))\text{-a.e.}$$

Thus, under the conditional independence assumption, the observation model can be simplified by dropping two arguments, the history and the action, from its domain. However, in the general case, the conditional independence assumption does not hold.

Some actions can be no-ops.

Example 4.1.5. Let $(\mu_n : H_n \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ be a schema. Consider the transition model $(\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$ and action $a \in A$ such that $\tau_n(h,a,y) = \delta_y$, for all $h \in H_{n-1}$, $y \in Y$, and $n \in \mathbb{N}$. Then, by Proposition A.2.7,

$$\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a, y) = \mu_n(h_n),$$

for all $h_n \in H_n$ and $n \in \mathbb{N}_0$. In other words, for an action having such a transition model, there is no change to the empirical belief during the transition update. That is, the action is a *no-op* (for that particular schema).

An observation model may be non-informative.

Example 4.1.6. Let $(\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$ be an observation model. Suppose that ξ satisfies the condition that, for all $h_n \in H_n$, $a_{n+1} \in A$, $o_{n+1} \in O$, and $n \in \mathbb{N}$,

$\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1}) : Y \to \mathbb{R}$ is a constant function, which may vary with h_n, a_{n+1}, o_{n+1}, and n. This is the case, for example, if O is a singleton set. Then

$$\begin{aligned}
&\mu_{n+1}(h_{n+1}) \\
&= \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1}) * (\mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y)) \\
&= \mu_n(h_n) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, y).
\end{aligned}$$

Thus ξ is an observation model that provides no information for the observation update, which therefore does not change the empirical belief produced by the transition update. Such an observation model is called *non-informative*.

Example 4.1.7. This example generalizes the hidden Markov model case of Example 2.3.4 to the case where the empirical belief has signature $\mathcal{P}(Y)$ and Y is finite. The underlying measure space is $(Y, 2^Y, c)$, where c is counting measure; thus υ_Y in Proposition 4.1.2 is c. Suppose that $Y = \{y_1, \ldots, y_m\}$. Then

$$\begin{aligned}
&\breve{\mu}_{n+1}(h_{n+1})(y_i) \\
&= \frac{\breve{\xi}_{n+1}(h_n, a_{n+1}, y_i)(o_{n+1}) \int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y_i)\, \breve{\mu}_n(h_n)\, dc}{\int_Y \left(\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1})\ \lambda s. \int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, y')(y)\, \breve{\mu}_n(h_n)\, dc\right) dc} \\
&\qquad\qquad \text{[Figure 4.5]} \\
&= \frac{\breve{\xi}_{n+1}(h_n, a_{n+1}, y_i)(o_{n+1}) \sum_{k=1}^m \breve{\tau}_{n+1}(h_n, a_{n+1}, y_k)(y_i)\, \breve{\mu}_n(h_n)(y_k)}{\sum_{j=1}^m (\breve{\xi}_{n+1}(h_n, a_{n+1}, y_j)(o_{n+1}) \sum_{k=1}^m \breve{\tau}_{n+1}(h_n, a_{n+1}, y_k)(y_j)\, \breve{\mu}_n(h_n)(y_k))}, \\
&\qquad\qquad \text{[Example A.2.6]}
\end{aligned}$$

for $i = 1, \ldots, m$. Consequently, $\breve{\mu}_{n+1}(h_{n+1})$ is a tractable expression (provided that m is not too large). Note that this result depends only on the finiteness of Y; there are no restrictions on A or O.

Example 4.1.8. Consider a schema of the form $(\mu_n : H_n \to \mathcal{P}(X \times Y))_{n \in \mathbb{N}_0}$. Then the transition model has components having signature

$$\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(X \times Y),$$

while the observation model has components having signature

$$\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O).$$

Now a component μ_n can be deconstructed into $\mu^{(1)} \otimes \mu^{(2)}$, where

$$\begin{aligned}
&\mu_n^{(1)} : H_n \to \mathcal{P}(X), \text{ and} \\
&\mu_n^{(2)} : H_n \times X \to \mathcal{P}(Y).
\end{aligned}$$

Consequently, one could expect the transition and observation models for $(\mu_n)_{n \in \mathbb{N}_0}$ to give a strong hint about what the transition and observation models for $(\mu_n^{(2)})_{n \in \mathbb{N}_0}$ should be. As shown in Section 4.2, the transition model for $(\mu_n^{(2)})_{n \in \mathbb{N}_0}$ is a variation of the one for $(\mu_n)_{n \in \mathbb{N}_0}$, and the observation models for $(\mu_n^{(2)})_{n \in \mathbb{N}_0}$ and $(\mu_n)_{n \in \mathbb{N}_0}$ are the same.

Here are some remarks to provide some intuition about observation models. Consider first the environment $(\Xi_n : H_{n-1} \times A \to \mathcal{P}(O))_{n\in\mathbb{N}}$. Thus, given $h_{n-1} \in H_{n-1}$ and $a_n \in A$, $\Xi_n(h_{n-1}, a_n)$ is a distribution for the possible values of o_n. The observation model $(\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$ introduces the extra argument Y into the domain of the environment. Having this extra argument generally provides more precise information about possible observations, in the sense that $\xi_n(h_{n-1}, a_n, y_n)$ 'sharpens' the distribution given by the environment about the possible values of o_n. The impact that the value of the Y argument has on the distribution of observations is exploited in the filtering process by using the observation model to compute the likelihood of the current observation that is then used in the observation update. (The function $\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, y)(o_{n+1}) : Y \to \mathbb{R}$ can be thought of as a likelihood function.)

The impact that Y has on the possible distributions on O can vary greatly. At one extreme, when Y is a state, the concomitant conditional independence assumptions imply that the observation model has the form $(\lambda(h, a, y).\xi'_n(y) : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$, where $(\xi'_n : Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$ is a sequence of regular conditional distributions. (See the proof of Proposition 2.3.2.) Essentially, in this case, the observation model does not depend on H_{n-1} nor A. This is the form of observation model considered in Chapter 2. However, generally at most one empirical belief in an agent's belief base will be a distribution on a state. Hence, for most empirical beliefs, all the domain arguments of the observation model will be needed. In fact, it will even be common for the Y argument to have only a small influence on the distribution of possible observations, so that the corresponding observation model will be similar to the environment where the Y argument is missing altogether. In such cases, the observation update will 'sharpen' the distribution that is empirical belief for Y only slightly. The extreme case is when

$$\sigma(\mathbf{y}_n) \underset{\sigma(\mathbf{h}_{n-1},\mathbf{a}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{o}_n),$$

for all $n \in \mathbb{N}$, so that

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)) = \mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n)) \text{ a.s.},$$

for all $B \in \mathcal{O}$, and the Y argument of the observation model can be dropped altogether.

Here is Bayes theorem for the nonconditional case. This result extends Proposition 2.3.3 beyond state schemas and is a special case of Proposition A.7.14.

Proposition 4.1.4. *(Bayes theorem for nonconditional schemas) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(Y, \mathcal{Y})$ standard Borel space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ a stochastic process, $(\Xi_n : H_{n-1} \times A \to \mathcal{P}(O))_{n\in\mathbb{N}}$ the environment for $\mathbf{a}$ and $\mathbf{o}$, $(\mu_n : H_n \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ the schema for $\mathbf{y}$, $(\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$ the transition model for $\mathbf{y}$, and $(\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$ the observation model for $\mathbf{y}$. Suppose that $\sigma(\mathbf{a}_{n+1}) \perp\!\!\!\perp_{\sigma(\mathbf{h}_n)} \sigma(\mathbf{y}_n)$, for all $n \in \mathbb{N}_0$. Then, for all $n \in \mathbb{N}_0$,*

$$\lambda(h, a).(\Xi_{n+1} \otimes \mu_{n+1})(h, a)(E^*) = \lambda(h, a).((\lambda(h, a).\mu_n(h) \odot \tau_{n+1}) \otimes \xi_{n+1})(h, a)(E)$$
$$\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))\text{-a.e.},$$

for all $E \in \mathcal{Y} \otimes \mathcal{O}$.

Proof. As shown in the proof of Proposition 4.1.2, for all $n \in \mathbb{N}_0$, $\lambda(h,a).\mu_n(h) \odot \tau_{n+1} : H_n \times A \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$. Also $\xi_{n+1} : H_n \times A \times Y \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{y}_{n+1})$. Hence, by Proposition A.7.12,

$$(\lambda(h,a).\mu_n(h) \odot \tau_{n+1}) \otimes \xi_{n+1} : H_n \times A \to \mathcal{P}(Y \times O)$$

is a regular conditional distribution of $(\mathbf{y}_{n+1}, \mathbf{o}_{n+1})$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

Now, for all $n \in \mathbb{N}$, $\Xi_{n+1} : H_n \times A \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$. Also $\mu_{n+1} : H_n \times A \times O \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{o}_{n+1})$. Hence, by Proposition A.7.12,

$$\Xi_{n+1} \otimes \mu_{n+1} : H_n \times A \to \mathcal{P}(O \times Y)$$

is a regular conditional distribution of $(\mathbf{o}_{n+1}, \mathbf{y}_{n+1})$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

The result now follows by the uniqueness part of Proposition A.5.16. □

The next result for this section provides a theoretical foundation for simulation in the nonconditional case.

Proposition 4.1.5. *(Simulation for the nonconditional case) Let $(A, \mathcal{A})$ be an action space, $(O, \mathcal{O})$ an observation space, $(Y, \mathcal{Y})$ a measurable space,*

$$\mu_0 : \mathcal{P}(Y)$$

a probability measure, and

$$(\Lambda_n : H_{n-1} \to \mathcal{P}(A))_{n \in \mathbb{N}},$$
$$(\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y))_{n \in \mathbb{N}},$$

and

$$(\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n \in \mathbb{N}}$$

sequences of probability kernels. Then there exists a probability space $(\Omega, \mathfrak{S}, \mathsf{P})$, an action process $\mathbf{a} : \Omega \to A^{\mathbb{N}}$, an observation process $\mathbf{o} : \Omega \to O^{\mathbb{N}}$, and a stochastic process $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ such that

$$\sigma(\mathbf{a}_n) \underset{\sigma(\mathbf{h}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}),$$
$$\sigma(\mathbf{y}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n),$$

and

$$\sigma(\mathbf{o}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{y}_n),$$

for all $n \in \mathbb{N}$. Furthermore,

$$(\Lambda_n : H_{n-1} \to \mathcal{P}(A))_{n \in \mathbb{N}}$$

is the agent for $\mathbf{a}$ *and* $\mathbf{o}$*,*

$$(\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$$

the transition model for $\mathbf{y}$*, and*

$$(\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$$

the observation model for $\mathbf{y}$*.*

Proof. Let $\Omega \triangleq Y \times A \times Y \times O \times A \times Y \times O \times \cdots$ and let $\mathfrak{S}$ be the usual product σ-algebra on Ω. Also let $\tilde{H}_n \triangleq Y \times (A \times Y \times O)^n$, for all $n \in \mathbb{N}_0$, and give each $\tilde{H}_n$ the usual product σ-algebra.

Define, for all $n \in \mathbb{N}$,

$$\begin{aligned}
\tilde{\Lambda}_n &\triangleq \lambda(y_0, a_1, y_1, o_1, \ldots, a_{n-1}, y_{n-1}, o_{n-1}).\Lambda_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}) : \tilde{H}_{n-1} \to \mathcal{P}(A)\\
\tilde{\tau}_n &\triangleq \lambda(y_0, a_1, y_1, o_1, \ldots, a_{n-1}, y_{n-1}, o_{n-1}, a_n).\tau_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}, a_n, y_{n-1})\\
&\qquad : \tilde{H}_{n-1} \times A \to \mathcal{P}(Y)\\
\tilde{\xi}_n &\triangleq \lambda(y_0, a_1, y_1, o_1, \ldots, a_{n-1}, y_{n-1}, o_{n-1}, a_n, y_n).\xi_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}, a_n, y_n)\\
&\qquad : \tilde{H}_{n-1} \times A \times Y \to \mathcal{P}(O).
\end{aligned}$$

Then each $\tilde{\Lambda}_n$, $\tilde{\tau}_n$, and $\tilde{\xi}_n$ is a probability kernel.

Define $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ to be the canonical projection. Hence $\mathbf{a}$ is an action process based on A. Also, define $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ to be the canonical projection. Hence $\mathbf{y}$ is a stochastic process. Similarly, define $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ to be the canonical projection. Hence $\mathbf{o}$ is an observation process based on O.

By Proposition A.8.1, there exists a unique probability measure P on $(\Omega, \mathfrak{S})$ such that, for all $n \in \mathbb{N}_0$,

$$\begin{aligned}
\mathsf{P} \circ (\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{y}_n, \mathbf{o}_n)^{-1} &= \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \otimes \tilde{\tau}_n \otimes \tilde{\xi}_n\\
\mathsf{P} \circ (\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{y}_n)^{-1} &= \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \otimes \tilde{\tau}_n\\
\mathsf{P} \circ (\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n)^{-1} &= \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n.
\end{aligned}$$

Thus $(\Omega, \mathfrak{S}, \mathsf{P})$ is a probability space.

Now it is shown that, for all $n \in \mathbb{N}$,

$$\sigma(\mathbf{a}_n) \underset{\sigma(\mathbf{h}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}).$$

Towards this, for all $n \in \mathbb{N}$ and $C \in \mathcal{A}$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}(\mathbf{a}_n^{-1}(C) \mid (\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}))\\
&= \lambda\omega.\tilde{\Lambda}_n((\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1})(\omega))(C) && \text{[Proposition A.8.2]}\\
&= \lambda\omega.\Lambda_n(\mathbf{h}_{n-1}(\omega))(C)\\
&= \mathsf{P}(\mathbf{a}_n^{-1}(C) \mid \mathbf{h}_{n-1}). && \text{[Proposition A.7.18]}
\end{aligned}$$

Hence the result. Furthermore, the first step above shows that each $\tilde{\Lambda}_n$ is a regular conditional distribution and the last step shows that Λ is the agent for $\mathbf{a}$ and $\mathbf{o}$.

Next it is shown that, for all $n \in \mathbb{N}$,

$$\sigma(\mathbf{y}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n).$$

Towards this, for all $n \in \mathbb{N}$ and $D \in \mathcal{Y}$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}(\mathbf{y}_n^{-1}(D) \mid (\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n)) \\
&= \lambda\omega.\tilde{\tau}_n((\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n)(\omega))(D) && \text{[Proposition A.8.2]} \\
&= \lambda\omega.\tau_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})(\omega))(D) \\
&= \mathsf{P}(\mathbf{y}_n^{-1}(D) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})). && \text{[Proposition A.7.18]}
\end{aligned}$$

Hence the result. Furthermore, the first step above shows that each $\tilde{\tau}_n$ is a regular conditional distribution and the last step shows that τ is the transition model for $\mathbf{y}$.

Also it is shown that, for all $n \in \mathbb{N}$,

$$\sigma(\mathbf{o}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{y}_n).$$

Towards this, for all $n \in \mathbb{N}$ and $E \in \mathcal{O}$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}(\mathbf{o}_n^{-1}(E) \mid (\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)) \\
&= \lambda\omega.\tilde{\xi}_n((\mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)(\omega))(E) && \text{[Proposition A.8.2]} \\
&= \lambda\omega.\xi_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)(\omega))(E) \\
&= \mathsf{P}(\mathbf{o}_n^{-1}(E) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_n)). && \text{[Proposition A.7.18]}
\end{aligned}$$

Hence the result. Furthermore, the first step above shows that each $\tilde{\xi}_n$ is a regular conditional distribution and the last step shows that ξ is the observation model for $\mathbf{y}$. □

The definitions of μ_0, Λ, τ, and ξ in Proposition 4.1.5 can be completely arbitrary. The result thus provides the basis for a simulation of an agent-environment system for which the choices of agent, initial empirical belief, transition model, and observation model are arbitrary.

To finish this section, a special case of significant interest for filtering is considered where the actual value in Y that the empirical belief is modelling is fixed over time. This happens, for example, in case Y is a parameter space and the parameter value is fixed over time and also in the case of Bayesian inference. The situation of the value in Y being fixed over time is captured by the requirement that the event $\{\omega \in \Omega \mid \mathbf{y}(\omega) \in Y^{\mathbb{N}_0}$ is a constant function (depending on $\omega)\}$ holds almost surely. This motivates the next definition.

Definition 4.1.3. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(Y, \mathcal{Y})$ a standard Borel space. A stochastic process $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is *constant-valued almost surely* if

$$\mathsf{P}(\{\omega \in \Omega \mid \mathbf{y}(\omega)(n) = \mathbf{y}(\omega)(n+1), \text{ for all } n \in \mathbb{N}_0\}) = 1.$$

In other words, $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is constant-valued almost surely if

$$\mathsf{P}(\{\omega \in \Omega \mid \mathbf{y}_n(\omega) = \mathbf{y}_{n+1}(\omega), \text{ for all } n \in \mathbb{N}_0\}) = 1.$$

Note that $\{\omega \in \Omega \mid \mathbf{y}_n(\omega) = \mathbf{y}_{n+1}(\omega), \text{ for all } n \in \mathbb{N}_0\}$ is measurable, that is, an event. To see this, consider the set $C \triangleq \{f \in Y^{\mathbb{N}_0} \mid f(n) = f(n+1), \text{ for all } n \in \mathbb{N}_0\}$, which is the set of functions in $Y^{\mathbb{N}_0}$ that are constant. According to Proposition A.4.2, C is measurable. (Let $\alpha_n : X \to X$ be the identity mapping, for all $n \in \mathbb{N}_0$, in Proposition A.4.2.) Since $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is measurable, $\mathbf{y}^{-1}(C)$ is a measurable subset of Ω. But $\{\omega \in \Omega \mid \mathbf{y}_n(\omega) = \mathbf{y}_{n+1}(\omega), \text{ for all } n \in \mathbb{N}_0\} = \mathbf{y}^{-1}(C)$.

'Constant-valued almost everywhere' is usually abbreviated to 'constant-valued a.s.'.

It is easy to show that $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is constant-valued a.s. if and only if $\mathbf{y}_n = \mathbf{y}_{n+1}$ a.s., for all $n \in \mathbb{N}_0$. In the case that $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is constant-valued a.s., the transition model τ has a particular form.

Proposition 4.1.6. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(Y, \mathcal{Y})$ a measurable space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ a stochastic process, and*

$$(\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y))_{n \in \mathbb{N}}$$

the transition model for $\mathbf{y}$. Suppose that $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is constant-valued a.s. Then, for all $n \in \mathbb{N}$,

$$\tau_n = \lambda(h, a, y).\delta_y \;\; \mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1}))\text{-a.e.}$$

Proof. For all $n \in \mathbb{N}$, since $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is constant-valued a.s.,

$$\int_\Omega \mathbf{1}_B \, \mathbf{1}_{\mathbf{y}_{n-1}^{-1}(C)} \, d\mathsf{P} = \int_\Omega \mathbf{1}_B \, \mathbf{1}_{\mathbf{y}_n^{-1}(C)} \, d\mathsf{P},$$

for all $B \in \sigma((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1}))$ and $C \in \mathcal{Y}$. Also, for all $n \in \mathbb{N}$, $\mathbf{1}_{\mathbf{y}_{n-1}^{-1}(C)}$ is $\sigma((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1}))$-measurable. It follows that $\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})) = \mathbf{1}_{\mathbf{y}_{n-1}^{-1}(C)}$ a.s, for all $C \in \mathcal{Y}$. Furthermore,

$$\lambda\omega.(\lambda(h, a, y).\delta_y)((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})(\omega))(C) = \lambda\omega.\delta_{\mathbf{y}_{n-1}(\omega)}(C) = \mathbf{1}_{\mathbf{y}_{n-1}^{-1}(C)}.$$

Thus, for all $n \in \mathbb{N}$,

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})) = \lambda\omega.(\lambda(h, a, y).\delta_y)((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{y}_{n-1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Hence the result. □

In other words, if $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is constant-valued a.s., then the transition model is such that every action is a no-op. This is consistent with the assumption that there are no actions and there is no transition model. In this case, a history is just a finite sequence of observations and filtering consists of a sequence of observation updates only. Also, in this setting, the observation model (for $\mathbf{y}$) is a sequence $\xi \triangleq (\xi_n)_{n \in \mathbb{N}}$, where

$$\xi_n : H_{n-1} \times Y \to \mathcal{P}(O)$$

is a regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{h}_{n-1}, \mathbf{y}_n)$, for all $n \in \mathbb{N}$.

Now the discussion turns to Bayesian inference in the nonconditional case. The next result shows that Bayesian inference in the nonconditional case is a special case of stochastic filtering. This result is important because it shows that all the methods of Bayesian machine learning can be regarded as special cases of stochastic filtering. The critical condition needed to reduce stochastic filtering to Bayesian inference is that $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ be constant-valued a.s.

Proposition 4.1.7. *(Bayesian inference in the nonconditional case) Let* $(\Omega, \mathfrak{S}, \mathsf{P})$ *be a probability space,* $(O, \mathcal{O})$ *an observation space,* $(Y, \mathcal{Y})$ *a standard Borel space,* $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ *an observation process,* $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ *a stochastic process,*

$$(\mu_n : H_n \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$$

the schema for $\mathbf{y}$*,*

$$(\xi_n : H_{n-1} \times Y \to \mathcal{P}(O))_{n \in \mathbb{N}}$$

the observation model for $\mathbf{y}$*,* υ_O *a* σ*-finite measure on* $\mathcal{O}$*, and* υ_Y *a* σ*-finite measure on* $\mathcal{Y}$*. Suppose that, for all* $n \in \mathbb{N}_0$*, there is a conditional density* $\breve{\mu}_n : H_n \to \mathcal{D}(Y)$ *such that* $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$ *and that, for all* $n \in \mathbb{N}$*, there is a conditional density* $\breve{\xi}_n : H_{n-1} \times Y \to \mathcal{D}(O)$ *such that* $\xi_n = \breve{\xi}_n \cdot \upsilon_O$*. Suppose also that* $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ *is constant-valued a.s. Then the following hold.*

1. *For all* $n \in \mathbb{N}_0$*,*

$$\mu_{n+1} = \lambda(h, o).\lambda y.\breve{\xi}_{n+1}(h, y)(o) * \lambda(h, o).\mu_n(h) \;\; \mathcal{L}(\mathbf{h}_{n+1})\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}(\mathbf{h}_{n+1})$*-almost all* $h_{n+1} \triangleq (h_n, o_{n+1}) \in H_{n+1}$*,*

$$\mu_{n+1}(h_{n+1}) = \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1}) * \mu_n(h_n).$$

3. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}(\mathbf{h}_{n+1})$*-almost all* $h_{n+1} \triangleq (h_n, o_{n+1}) \in H_{n+1}$*,*

$$\breve{\mu}_{n+1}(h_{n+1}) = \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1}) * \breve{\mu}_n(h_n) \quad \upsilon_Y\text{-a.e.}$$

Proof. 1. For all $n \in \mathbb{N}_0$, $\mu_n : H_n \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $\mathbf{h}_n$ and so

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid \mathbf{h}_n) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Since $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ is constant-valued a.s., it follows that $\mathbf{1}_{\mathbf{y}_n^{-1}(C)} = \mathbf{1}_{\mathbf{y}_{n+1}^{-1}(C)}$ a.s., for all $C \in \mathcal{Y}$, and so

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid \mathbf{h}_n) = \mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid \mathbf{h}_n) \text{ a.s.}$$

Thus

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid \mathbf{h}_n) = \lambda\omega.\mu_n(\mathbf{h}_n(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. That is, for all $n \in \mathbb{N}_0$, $\mu_n : H_n \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $\mathbf{h}_n$.

Next, ξ_{n+1} is an observation model; hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{y}_{n+1})) = \lambda\omega.\xi_{n+1}((\mathbf{h}_n, \mathbf{y}_{n+1})(\omega))(B) \text{ a.s.}$$

Also $\breve{\xi}_{n+1}$ is a regular conditional density.

Now consider the probability kernel

$$\lambda(h, o).\lambda y.\breve{\xi}_{n+1}(h, y)(o) * \lambda(h, o).\mu_n(h) : H_n \times O \to \mathcal{P}(Y).$$

By Proposition A.12.8, this probability kernel is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{o}_{n+1})$. That is,

$$\lambda(h, o).\lambda y.\breve{\xi}_{n+1}(h, y)(o) * \lambda(h, o).\mu_n(h) : H_{n+1} \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $\mathbf{h}_{n+1}$. Since μ_{n+1} is by definition a regular conditional distribution of $\mathbf{y}_{n+1}$ given $\mathbf{h}_{n+1}$, it follows from the uniqueness part of Proposition A.5.16 that

$$\mu_{n+1} = \lambda(h, o).\lambda y.\breve{\xi}_{n+1}(h, y)(o) * \lambda(h, o).\mu_n(h) \;\; \mathcal{L}(\mathbf{h}_{n+1})\text{-a.e.}$$

2. Hence, for all $n \in \mathbb{N}$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, o_{n+1}) \in H_{n+1}$,

$$\mu_{n+1}(h_{n+1}) = \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1}) * \mu_n(h_n).$$

3. For all $n \in \mathbb{N}$ and $\mathcal{L}(\mathbf{h}_{n+1})$-almost all $h_{n+1} \triangleq (h_n, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\breve{\mu}_{n+1}(h_{n+1}) \cdot \upsilon_Y \\
&= \mu_{n+1}(h_{n+1}) \\
&= \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1}) * \mu_n(h_n) \\
&= \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1}) * (\breve{\mu}_n(h_n) \cdot \upsilon_Y) \\
&= (\lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1}) * \breve{\mu}_n(h_n)) \cdot \upsilon_Y. \qquad \text{[Proposition A.3.10]}
\end{aligned}$$

The result now follows by Proposition A.2.11. □

$$\mu_{n+1} = \lambda(h, o).\lambda y.\breve{\xi}_{n+1}(h, y)(o) * \lambda(h, o).\mu_n(h)$$

$$\mu_{n+1}(h_{n+1}) = \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1}) * \mu_n(h_n)$$

$$\breve{\mu}_{n+1}(h_{n+1}) = \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1}) * \breve{\mu}_n(h_n)$$

Figure 4.6: Bayesian inference in the nonconditional case

The recurrence equations for Bayesian inference are given in Figures 4.6 and 4.7. They show that stochastic filtering is a generalization of Bayesian inference.

$$\mu_{n+1}(h_{n+1}) = \lambda B.\frac{\int_Y \mathbf{1}_B\, \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1})\; d\mu_n(h_n)}{\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1})\; d\mu_n(h_n)}$$

$$\breve{\mu}_{n+1}(h_{n+1}) = \frac{\lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1})\; \breve{\mu}_n(h_n)}{\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, y)(o_{n+1})\; \breve{\mu}_n(h_n)\; d\upsilon_Y}$$

Figure 4.7: Explicit form of Bayesian inference in the nonconditional case

4.2 Conditional Filters

To begin, here is some motivation for the material of this section. Consider a schema having the form

$$(\mu_n : H_n \to \mathcal{P}(X \times Y))_{n\in\mathbb{N}_0}.$$

Assuming that the modelling roles of X and Y are similar, this schema should be treated as in Section 4.1 with $X \times Y$ replacing the Y there. Thus the transition model should have the form

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(X \times Y))_{n\in\mathbb{N}_0}$$

and the observation model should have the form

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}_0}.$$

Filtering for this case proceeds exactly as in Section 4.1 with $X \times Y$ replacing the Y there.

Another possibility is that the modelling roles of X and Y are different. A typical such case is when X is a parameter space and the value of the parameter is fixed but unknown, and the primary interest is to track the distribution on Y corresponding to the actual value of the parameter. In this situation, as shall become clear below, it is advantageous to deconstruct the schema into

$$(\mu^{(1)}_n : H_n \to \mathcal{P}(X))_{n\in\mathbb{N}_0}$$

and

$$(\mu^{(2)}_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}.$$

Filtering steps are done successively on $\mu^{(1)}$ and then on $\mu^{(2)}$. The filtering output is a distribution on the possible value of the parameter in X and a conditional distribution on Y given X. In effect, at each filtering step, the parameter value in X is estimated first and then the corresponding distribution on Y is tracked. For this asymmetric handling of X and Y to work, it is necessary to place a restriction on the stochastic process $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$; in particular, there should be a functional relationship between $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$. The strong but nevertheless highly useful assumption that the stochastic process $\mathbf{x}$ be constant-valued

a.s. is made in Section 4.2.1. A weaker assumption that however slightly complicates the filter recurrence equations will be employed in Section 4.2.2.

Thus, with the above motivation, the setting for this section is that of schemas having the form

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}.$$

Compared with the schemas of Section 4.1, the difference is in the addition of the argument X in the domain. This means that the transaction and observation models of Section 4.1 must be modified in an analogous way. The setting of Section 4.1 versus the schemas, transition models, and observation models of this section that result from this modification is summarized in Figure 4.8.

$\mu_n : H_n \to \mathcal{P}(Y)$	$\mu_n : H_n \times X \to \mathcal{P}(Y)$
$\tau_n : H_{n-1} \times A \times Y \to \mathcal{P}(Y)$	$\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y)$
$\xi_n : H_{n-1} \times A \times Y \to \mathcal{P}(O)$	$\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O)$

Figure 4.8: Comparison of transition models and observation models for the schemas of Section 4.1 versus the schemas of this section

Note for later use that, if $(\nu_n : H_n \to \mathcal{P}(X))_{n\in\mathbb{N}_0}$ is a schema for $\mathbf{x}$ and $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ is a schema for $\mathbf{y}$ given $\mathbf{x}$, then, for all $n \in \mathbb{N}_0$, $\nu_n \odot \mu_n : H_n \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $\mathbf{h}_n$, that is,

$$\mathsf{P}(\mathbf{y}_n^{-1}(B) \mid \mathbf{h}_n) = \lambda\omega.(\nu_n \odot \mu_n)(\mathbf{h}_n(\omega))(B)) \text{ a.s.},$$

for all $B \in \mathcal{Y}$. Also $\nu_n \otimes \mu_n : H_n \to \mathcal{P}(X \times Y)$ is a regular conditional distribution of $(\mathbf{x}_n, \mathbf{y}_n)$ given $\mathbf{h}_n$, that is,

$$\mathsf{P}((\mathbf{x}_n, \mathbf{y}_n)^{-1}(C) \mid \mathbf{h}_n) = \lambda\omega.(\nu_n \otimes \mu_n)(\mathbf{h}_n(\omega))(C)) \text{ a.s.},$$

for all $C \in \mathcal{X} \otimes \mathcal{Y}$.

Here are the definitions of the transition and observation models for the setting of this section. In essence, the extra domain argument X that distinguishes the signature $H_n \times X \to \mathcal{P}(Y)$ from $H_n \to \mathcal{P}(Y)$ also appears as a domain argument in the transition model and observation model for $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$.

Definition 4.2.1. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ measurable spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes. A *transition model* (*for* $\mathbf{y}$ *given* $\mathbf{x}$) is a sequence $\tau \triangleq (\tau_n)_{n\in\mathbb{N}}$, where

$$\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})$, for all $n \in \mathbb{N}$.

In other words, for all $n \in \mathbb{N}$, τ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})) = \lambda\omega.\tau_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})(\omega))(C) \text{ a.s.},$$

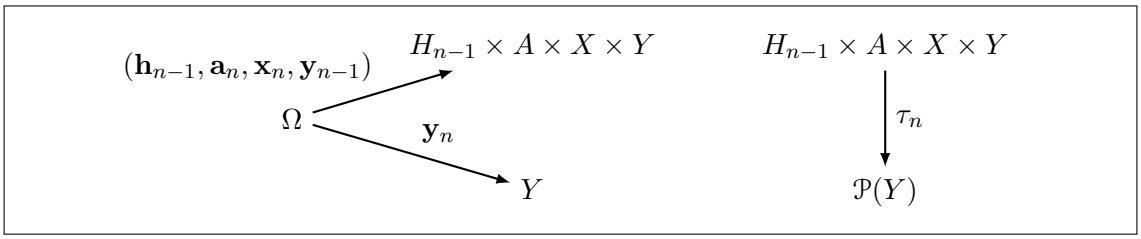

Figure 4.9: A component of a transition model

for all $C \in \mathcal{Y}$. According to Proposition A.5.16, assuming that Y is a standard Borel space, for each $n \in \mathbb{N}$, such a τ_n exists and is unique $\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}))$-a.e.

A component of a transition model takes as input a value in $H_{n-1} \times A \times X \times Y$ and returns a distribution on the values in Y that could result from the transition. Here, H_{n-1} is the set of histories up to the current observation received by the agent. Having this extra argument is a requirement that comes from replacing the state space S by an arbitrary set Y; elements of Y (or even $X \times Y$) may not force the conditional independence properties that states do.

Now comes the definition of an observation model.

Definition 4.2.2. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ measurable spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes. An *observation model* (*for* $\mathbf{y}$ *given* $\mathbf{x}$) is a sequence $\xi \triangleq (\xi_n)_{n \in \mathbb{N}}$, where

$$\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O)$$

is a regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)$, for all $n \in \mathbb{N}$.

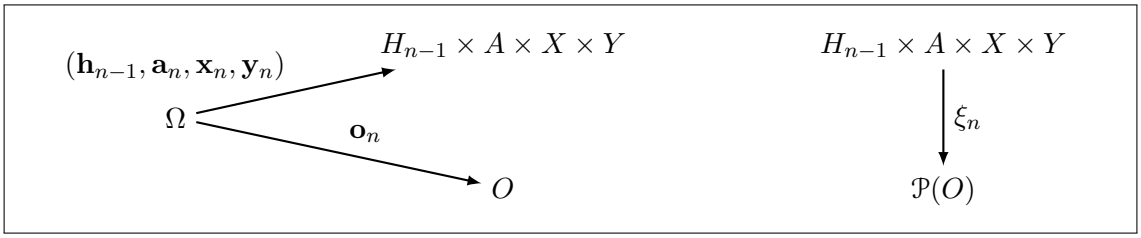

Figure 4.10: A component of an observation model

In other words, for all $n \in \mathbb{N}$, ξ_n is a probability kernel that satisfies the condition

$$\mathsf{P}(\mathbf{o}_n^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)) = \lambda\omega.\xi_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{O}$. According to Proposition A.5.16, for each $n \in \mathbb{N}$, such a ξ_n exists and is unique $\mathcal{L}((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n))$-a.e.

A component of an observation model takes as input a value in $H_{n-1} \times A \times X \times Y$ and returns a distribution on the observations that could be received by the agent.

It is noteworthy that *each schema can have its own observation space* different from the observations spaces of other schemas. This also means that the history can be different for

different schemas. However, when empirical beliefs are obtained by instantiating schemas with the current history for those schemas, this distinction disappears since the history is no longer part of the domain. This flexibility of being able to vary the observation space is important in practice. Some applications will have tens or hundreds or even thousands of empirical beliefs (not just a single state distribution) and the kinds of observations that are most effective for learning these empirical beliefs may vary. Perhaps the simplest way to think about this is to imagine that there is a single large observation that is available to the agent at each time step and, depending on the particular empirical belief, a certain part of this large observation is relevant and effective for filtering that empirical belief. Note also that the range under a specific function of the large observation can also be used as an observation; thus features of observations can also be used as observations.

Some reflection will show that, in the setting of the section, the following conditional independence property is a reasonable assumption:

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$. This property cannot be *proved* without specific assumptions about the dependency graph for the schema, indeed, for the highly complex graph which underlies all the schemas in the schema base taken together. Because the strong assumptions for state schemas are not available in this more general context, this graph has plenty of edges between typical pairs of nodes, making conditional independence properties hard to find. But it seems that for the particular property above, one can reasonably expect that $\{\mathbf{a}_{n+1}\}$ and $\{\mathbf{y}_n\}$ will be d-separated by $\{\mathbf{h}_n, \mathbf{x}_n\}$, and hence the property will hold. Furthermore, in the context of a simulation for the setting of this section, the conditional independence property can be be *proved* to hold. (See the proof of Proposition 4.2.7 below.)

Here is a consequence of the conditional independence assumption.

Proposition 4.2.1. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes, $n \in \mathbb{N}_0$, and $\mu_n : H_n \times X \to \mathcal{P}(Y)$ a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{x}_n)$ Suppose that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n).$$

Then $\lambda(h, a, x).\mu_n(h, x) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)$.

Proof. Note that $\lambda(h, a, x).\mu_n(h, x)$ is measurable, that is, a probability kernel. Since μ_n is a regular conditional distribution,

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{x}_n)) = \lambda\omega.\mu_n((\mathbf{h}_n, \mathbf{x}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Now, since

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

Proposition A.6.1 shows that

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{x}_n)) = \mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Thus

$$\mathsf{P}((\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)) = \lambda\omega.(\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. In other words, $\lambda(h,a,x).\mu_n(h,x) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)$. □

4.2.1 Constant-valued Case

The setting of this subsection is that the value in X is unknown but fixed. This property is captured by the requirement that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ be constant-valued a.s. (See Definition 4.1.3. The next subsection is concerned with a less restrictive requirement that, however, requires a complication of the filter recurrence equations.) Here is a consequence of the constant-valued a.s. assumption.

Proposition 4.2.2. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes, $n \in \mathbb{N}_0$, and $\mu_n : H_n \times X \to \mathcal{P}(Y)$ a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{x}_n)$. Suppose that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

and also that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. Then $\lambda(h,a,x).\mu_n(h,x) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Proof. By Proposition 4.2.1,

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)) = \lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Since $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s., it follows that

$$\lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)(\omega))(C) = \lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Next, let $D \in \mathcal{H}_n \otimes \mathcal{A} \otimes \mathcal{X}$. Let $B \triangleq (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)^{-1}(D)$, so that $B \in \sigma((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n))$, and $B' \triangleq (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})^{-1}(D)$, so that $B' \in \sigma((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$. Since $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s., $\mathbf{1}_B = \mathbf{1}_{B'}$ a.s. Then, for all $C \in \mathcal{Y}$,

$$\begin{aligned}
&\int_\Omega \mathbf{1}_{B'}\, \mathbf{1}_{\mathbf{y}_n^{-1}(C)}\, d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_B\, \mathbf{1}_{\mathbf{y}_n^{-1}(C)}\, d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_B\, \lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)(\omega))(C)\, d\mathsf{P} \\
&\qquad\qquad [\lambda(h,a,x).\mu_n(h,x) \text{ is a regular conditional distribution}] \\
&= \int_\Omega \mathbf{1}_{B'}\, \lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C)\, d\mathsf{P}.
\end{aligned}$$

Also $\lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n,\mathbf{a}_{n+1},\mathbf{x}_{n+1})(\omega))(C)$ is $\sigma((\mathbf{h}_n,\mathbf{a}_{n+1},\mathbf{x}_{n+1}))$-measurable, for all $C \in \mathcal{Y}$. Thus

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n,\mathbf{a}_{n+1},\mathbf{x}_{n+1})) = \lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n,\mathbf{a}_{n+1},\mathbf{x}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. That is, $\lambda(h,a,x).\mu_n(h,x) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n,\mathbf{a}_{n+1},\mathbf{x}_{n+1})$. □

Now comes the result giving the filter recurrence equations for schemas and empirical beliefs in the conditional case, assuming that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s.

Proposition 4.2.3. *(Filter recurrence equations for the constant-valued, conditional case) Let $(\Omega,\mathfrak{S},\mathsf{P})$ be a probability space, $(A,\mathcal{A})$ an action space, $(O,\mathcal{O})$ an observation space, $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes,*

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$$

the schema for $\mathbf{y}$ given $\mathbf{x}$,

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$$

the transition model for $\mathbf{y}$ given $\mathbf{x}$,

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$$

the observation model for $\mathbf{y}$ given $\mathbf{x}$, υ_O a σ-finite measure on $\mathcal{O}$, υ_X a σ-finite measure on $\mathcal{X}$, and υ_Y a σ-finite measure on $\mathcal{Y}$. Suppose that, for all $n \in \mathbb{N}_0$, there is a conditional density $\breve{\mu}_n : H_n \times X \to \mathcal{D}(Y)$ such that $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$ and that, for all $n \in \mathbb{N}$, there is a conditional density $\breve{\tau}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(Y)$ such that $\tau_n = \breve{\tau}_n \cdot \upsilon_Y$ and a conditional density $\breve{\xi}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(O)$ such that $\xi_n = \breve{\xi}_n \cdot \upsilon_O$. Suppose also that

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n,\mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$, and that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. Then the following hold.

1. *For all $n \in \mathbb{N}_0$,*

$$\mu_{n+1} = \lambda(h,a,o,x).\lambda y.\breve{\xi}_{n+1}(h,a,x,y)(o) * \lambda(h,a,o,x).(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})(h,a,x) \quad \mathcal{L}((\mathbf{h}_{n+1},\mathbf{x}_{n+1}))\text{-a.e.}$$

2. *For all $n \in \mathbb{N}_0$ and $h_{n+1} \triangleq (h_n,a_{n+1},o_{n+1}) \in H_{n+1}$,*

$$\lambda x.\mu_{n+1}(h_{n+1},x) = \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * \lambda x.(\mu_n(h_n,x) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},x,y)),$$

except on $\{x \in X \mid (h_{n+1},x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1},\mathbf{x}_{n+1}))(N_{n+1}) = 0$.

3. For all $n \in \mathbb{N}_0$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}&\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) = \\ &\lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) \;*\; \lambda x.(\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \upsilon_Y\text{-a.e.},\end{aligned}$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$.

Proof. 1. Clearly, $\lambda(h, a, x).\breve{\mu}_n(h, x) : H_n \times A \times X \to \mathcal{D}(Y)$ is a conditional density, for all $n \in \mathbb{N}_0$. Also, for all $n \in \mathbb{N}_0$,

$$\lambda(h, a, x).\mu_n(h, x) = \lambda(h, a, x).\breve{\mu}_n(h, x) \cdot \upsilon_Y.$$

To see this, since $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$, it follows that

$$\begin{aligned}&(\lambda(h, a, x).\breve{\mu}_n(h, x) \cdot \upsilon_Y)(h, a, x)(C) \\ &= \int_Y \mathbf{1}_C \, \lambda(h, a, x).\breve{\mu}_n(h, x)(h, a, x) \, d\upsilon_Y \\ &= \int_Y \mathbf{1}_C \, \breve{\mu}_n(h, x) \, d\upsilon_Y \\ &= (\breve{\mu}_n \cdot \upsilon_Y)(h, x)(C) \\ &= \mu_n(h, x)(C) \\ &= \lambda(h, a, x).\mu_n(h, x)(h, a, x)(C),\end{aligned}$$

for all $(h, a, x) \in H_n \times A \times X$ and $C \in \mathcal{Y}$. Hence $\lambda(h, a, x).\mu_n(h, x) = \lambda(h, a, x).\breve{\mu}_n(h, x) \cdot \upsilon_Y$.

By Proposition 4.2.2, for all $n \in \mathbb{N}_0$, $\lambda(h, a, x).\mu_n(h, x) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Now, for all $n \in \mathbb{N}_0$, since τ_{n+1} is a regular conditional distribution,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)) = \lambda\omega.\tau_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Proposition A.7.15 now shows that, for all $C \in \mathcal{Y}$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})) = \lambda\omega.(\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1})((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.}$$

That is, for all $n \in \mathbb{N}_0$, $\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1} : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Furthermore,

$$\begin{aligned}&\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1} \\ &= (\lambda(h, a, x).\breve{\mu}_n(h, x) \cdot \upsilon_Y) \odot (\breve{\tau}_{n+1} \cdot \upsilon_Y) \\ &= (\lambda(h, a, x).\breve{\mu}_n(h, x) \odot \breve{\tau}_{n+1}) \cdot \upsilon_Y. \qquad \text{[Proposition A.3.8]}\end{aligned}$$

Hence $\lambda(h, a, x).\breve{\mu}_n(h, x) \odot \breve{\tau}_{n+1}$ is a regular conditional density of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Next, ξ_{n+1} is an observation model; hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})) = \lambda\omega.\xi_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})(\omega))(B).$$

Also $\breve{\xi}_{n+1}$ is a regular conditional density.

Now consider the probability kernel

$$\lambda(h,a,o,x).\lambda y.\breve{\xi}_{n+1}(h,a,x,y)(o) \;*\; \lambda(h,a,o,x).(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})(h,a,x) : \\ H_n \times A \times O \times X \to \mathcal{P}(Y).$$

By Proposition A.12.8, this probability kernel is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{o}_{n+1}, \mathbf{x}_{n+1})$. Hence

$$\lambda(h,a,o,x).\lambda y.\breve{\xi}_{n+1}(h,a,x,y)(o) * \lambda(h,a,o,x).(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})(h,a,x) : \\ H_{n+1} \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_{n+1}, \mathbf{x}_{n+1})$. Since μ_{n+1} is by definition a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_{n+1}, \mathbf{x}_{n+1})$, it follows from the uniqueness part of Proposition A.5.16 that, for all $n \in \mathbb{N}_0$,

$$\mu_{n+1} = \\ \lambda(h,a,o,x).\lambda y.\breve{\xi}_{n+1}(h,a,x,y)(o) * \lambda(h,a,o,x).(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})(h,a,x) \\ \mathcal{L}((\mathbf{h}_{n+1}, \mathbf{h}_{n+1}))\text{-a.e.}$$

2. Hence, for all $n \in \mathbb{N}$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned} &\lambda x.\mu_{n+1}(h_{n+1},x) \\ &= \lambda x.(\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * (\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})(h_n,a_{n+1},x)) \\ &= \lambda x.(\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * (\mu_n(h_n,x) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},x,y))) \\ &= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * \lambda x.(\mu_n(h_n,x) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},x,y)), \end{aligned}$$

except on $\{x \in X \mid (h_{n+1},x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$.

3. For this part, recall Definition A.3.6. For all $n \in \mathbb{N}$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned} &\lambda x.\breve{\mu}_{n+1}(h_{n+1},x) \cdot \upsilon_Y \\ &= \lambda x.\mu_{n+1}(h_{n+1},x) \\ &= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * \lambda x.(\mu_n(h_n,x) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},x,y)) \\ &= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * \lambda x.((\breve{\mu}_n(h_n,x) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},x,y) \cdot \upsilon_Y)) \\ &= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * \lambda x.((\breve{\mu}_n(h_n,x) \odot \lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},x,y)) \cdot \upsilon_Y) \\ &\qquad \text{[Proposition A.3.8]} \\ &= (\lambda x.\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * \lambda x.(\breve{\mu}_n(h_n,x) \odot \lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},x,y))) \cdot \upsilon_Y. \\ &\qquad \text{[Proposition A.3.10]} \end{aligned}$$

except on $\{x \in X \mid (h_{n+1},x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$. The result now follows by Proposition A.2.11. □

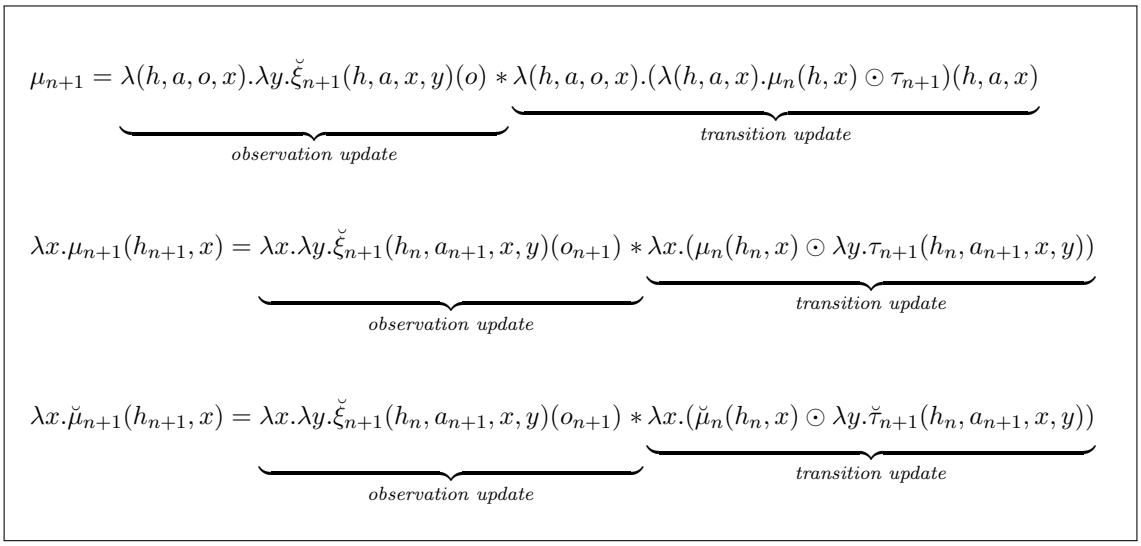

Figure 4.11: Recurrence equations for schemas and empirical beliefs

Given the initial empirical belief $\lambda x.\mu_0(h_0,x)$, the recurrence equation in Part 2 of Proposition 4.2.3 enables the computation of $\lambda x.\mu_1(h_1,x)$, $\lambda x.\mu_2(h_2,x)$, and so on, where $h_0 \triangleq (), h_1, h_2, \ldots$ are the successive histories.

Proposition 4.1.2 can be obtained as a special case of Proposition 4.2.3 by letting X be a singleton set.

The assumption that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. plays a crucial role in the proof of Proposition 4.2.3. Observe that $\lambda(h,a,x).\mu_n(h,x)$ is naturally a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)$. However, to form the fusion of $\lambda(h,a,x).\mu_n(h,x)$ with τ_{n+1}, one needs a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$. The simplest way to achieve this is to assume that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. More generally, it is necessary to have some way of materializing a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$ from a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)$.

Since the observation space O is typically structured, the space $\mathcal{P}(O)$ in the observation model is a space of probabilities over a structured space. However, there is never any need to deconstruct the observation model in the way that schemas are deconstructed. For a suitable function

$$\overline{f}_{\xi_n} : H_{n-1} \times A \to X \to Y \to \mathcal{D}(O),$$

one can define

$$\breve{\xi}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(O)$$

by

$$\breve{\xi}_n = \lambda(h,a,x,y).\overline{f}_{\xi_n}(h,a)(x)(y).$$

Thus, for all $h_n \in H_n$ and $a_{n+1} \in A$,

$$\overline{f}_{\xi_{n+1}}(h_n,a_{n+1}) : X \to Y \to \mathcal{D}(O)$$

can be neatly extracted. Then, for use in Parts 2 and 3 of Proposition 4.2.3,

$$\lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) = \lambda x.\lambda y.\overline{f}_{\xi_{n+1}}(h_n, a_{n+1})(x)(y)(o_{n+1}).$$

Using the definition of the projective product, the recurrence equation for empirical beliefs (in probability measure form) is more explicitly as follows: for all $n \in \mathbb{N}_0$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\lambda x.\mu_{n+1}(h_{n+1}, x) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) \ * \ \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \\
&= \lambda x.\lambda B.\frac{\int_Y \mathbf{1}_B \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ d(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))}{\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ d(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))} \\
&= \lambda x.\lambda B.\frac{\int_Y \left(\lambda y'.\int_Y \mathbf{1}_B \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ d\tau_{n+1}(h_n, a_{n+1}, x, y')\right)\ d\mu_n(h_n, x)}{\int_Y \left(\lambda y'.\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ d\tau_{n+1}(h_n, a_{n+1}, x, y')\right)\ d\mu_n(h_n, x)},
\end{aligned}$$

[Proposition A.7.8]

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$. Thus, for all $n \in \mathbb{N}_0$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\lambda x.\mu_{n+1}(h_{n+1}, x) = \\
&\lambda x.\lambda B.\frac{\int_Y \left(\lambda y'.\int_Y \mathbf{1}_B \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ d\tau_{n+1}(h_n, a_{n+1}, x, y')\right)\ d\mu_n(h_n, x)}{\int_Y \left(\lambda y'.\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ d\tau_{n+1}(h_n, a_{n+1}, x, y')\right)\ d\mu_n(h_n, x)},
\end{aligned}$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$.

Similarly, the recurrence equation for empirical beliefs in density form is more explicitly as follows: for all $n \in \mathbb{N}_0$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$, υ_Y-almost everywhere,

$$\begin{aligned}
&\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) * \lambda x.(\breve{\mu}_n(h_n, x) \ \odot \ \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \\
&= \lambda x.\frac{\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ (\breve{\mu}_n(h_n, x) \ \odot \ \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))}{\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ (\breve{\mu}_n(h_n, x) \ \odot \ \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))\ d\upsilon_Y} \\
&= \lambda x.\frac{\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y')(y)\ \breve{\mu}_n(h_n, x)\ d\upsilon_Y}{\int_Y(\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y')(y)\ \breve{\mu}_n(h_n, x)\ d\upsilon_Y)\ d\upsilon_Y},
\end{aligned}$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$.

Thus, for all $n \in \mathbb{N}_0$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$, υ_Y-almost everywhere,

$$\begin{aligned}
&\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) = \\
&\lambda x.\frac{\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y')(y)\ \breve{\mu}_n(h_n, x)\ d\upsilon_Y}{\int_Y(\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y')(y)\ \breve{\mu}_n(h_n, x)\ d\upsilon_Y)\ d\upsilon_Y},
\end{aligned}$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$.

The two preceding recurrence equations in explicit form are illustrated in Figure 4.12.

$$\lambda x.\mu_{n+1}(h_{n+1}, x) =$$
$$\lambda x.\lambda B.\frac{\int_Y \left(\lambda y'.\int_Y \mathbf{1}_B \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ d\tau_{n+1}(h_n, a_{n+1}, x, y')\right)\ d\mu_n(h_n, x)}{\int_Y \left(\lambda y'.\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ d\tau_{n+1}(h_n, a_{n+1}, x, y')\right)\ d\mu_n(h_n, x)}$$

$$\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) =$$
$$\lambda x.\frac{\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y')(y)\ \breve{\mu}_n(h_n, x)\ d\upsilon_Y}{\int_Y (\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1})\ \lambda y.\int_Y \lambda y'.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y')(y)\ \breve{\mu}_n(h_n, x)\ d\upsilon_Y)\ d\upsilon_Y}$$

Figure 4.12: Recurrence equations for empirical beliefs in explicit form

From a theoretical point of view, Proposition 4.2.3 provides a straightforward account of filtering. However, in practice, there is a major problem: except in a few special cases, the expression for the updated empirical belief is not tractable. This means that, while there is an explicit mathematical expression for the updated empirical belief, it is not easily calculable. The problem is that, except in a few cases, the result of the transition update and the result of the observation update cannot be simplified, so that the syntactic size of the expression for the updated empirical belief is linear in the number of time steps. After even a small number of time steps, the expression becomes intractable. There are two important cases where simplification is possible and the resulting expression is tractable: one is where the state distribution is Gaussian and the transaction and observation models are linear Gaussian (the case of linear dynamical systems) and the other is where the state space is finite and the state distribution is categorical (the case of hidden Markov models). More generally, the expression for the updated empirical belief is usually not tractable and a different approach is required. The standard solution for state distributions is to use a particle filter. For this, instead of maintaining an explicit mathematical expression for a state distribution, a suitable collection of states (the 'particles') is maintained that approximates the state distribution. Then, for example, the integral with respect to the state distribution of a utility function can be computed by standard Monte Carlo methods. This approach has been so successful that particle filters are used almost universally in robotics and autonomous vehicle applications, and it is taken up in detail in Section 4.3.

From the ingredients of Propositions 4.1.2 and 4.2.3, the environment can be synthesized. In this section, it will be convenient to use the notation $(\nu_n : H_n \to \mathcal{P}(X))_{n\in\mathbb{N}_0}$, $(\eta_n : H_{n-1} \times A \times X \to \mathcal{P}(X))_{n\in\mathbb{N}}$, and $(\zeta_n : H_{n-1} \times A \times X \to \mathcal{P}(O))_{n\in\mathbb{N}}$ for the schema, transition model, and observation model, respectively, from Section 4.1 (instead of μ, τ, and ξ, respectively).

Proposition 4.2.4. *(Environment synthesis for the constant-valued, conditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an*

observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ *and* $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ *stochastic processes,*

$$(\nu_n : H_n \to \mathcal{P}(X))_{n\in\mathbb{N}_0}$$

the schema for $\mathbf{x}$*,*

$$(\eta_n : H_{n-1} \times A \times X \to \mathcal{P}(X))_{n\in\mathbb{N}}$$

the transition model for $\mathbf{x}$*,*

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$$

the schema for $\mathbf{y}$ *given* $\mathbf{x}$*,*

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$$

the transition model for $\mathbf{y}$ *given* $\mathbf{x}$*,*

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$$

the observation model for $\mathbf{y}$ *given* $\mathbf{x}$*,* υ_O *a* σ*-finite measure on* $\mathcal{O}$*,* υ_X *a* σ*-finite measure on* $\mathcal{X}$*, and* υ_Y *a* σ*-finite measure on* $\mathcal{Y}$*. Let* $(\Xi_n : H_{n-1}\times A \to \mathcal{P}(O))_{n\in\mathbb{N}}$ *be the environment for* $\mathbf{a}$ *and* $\mathbf{o}$*. Suppose that, for all* $n \in \mathbb{N}_0$*, there is a conditional density* $\breve{\nu}_n : H_n \to \mathcal{D}(X)$ *such that* $\nu_n = \breve{\nu}_n \cdot \upsilon_X$*, and a conditional density* $\breve{\mu}_n : H_n \times X \to \mathcal{D}(Y)$ *such that* $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$*. Suppose that, for all* $n \in \mathbb{N}$*, there is a conditional density* $\breve{\Xi}_n : H_{n-1}\times A \to \mathcal{D}(O)$ *such that* $\Xi_n = \breve{\Xi}_n \cdot \upsilon_O$*, a conditional density* $\breve{\eta}_n : H_{n-1} \times A \times X \to \mathcal{D}(X)$ *such that* $\eta_n = \breve{\eta}_n \cdot \upsilon_X$*, a conditional density* $\breve{\tau}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(Y)$ *such that* $\tau_n = \breve{\tau}_n \cdot \upsilon_Y$*, and a conditional density* $\breve{\xi}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(O)$ *such that* $\xi_n = \breve{\xi}_n \cdot \upsilon_O$*. Suppose also that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{x}_n)$$

and

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n,\mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$*, and that* $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ *is constant-valued a.s. Then the following hold.*

1. *For all* $n \in \mathbb{N}_0$*,*

$$\Xi_{n+1} = \\ ((\lambda(h,a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})) \odot \xi_{n+1} \\ \mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$*,*

$$\Xi_{n+1}(h_n, a_{n+1}) = \\ ((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \odot \\ \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y).$$

3. *For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$,*

$$\breve{\Xi}_{n+1}(h_n, a_{n+1}) = \\ ((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \odot \\ \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \quad \upsilon_O\text{-a.e.}$$

4. *For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$, and all $C \in \mathcal{O}$,*

$$\Xi_{n+1}(h_n, a_{n+1})(C) = \\ \int_{X \times Y} \lambda(x, y).\xi_{n+1}(h_n, a_{n+1}, x, y)(C) \\ d((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))).$$

5. *For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$, and υ_O-almost all $o \in O$,*

$$\breve{\Xi}_{n+1}(h_n, a_{n+1})(o) = \\ \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\ d((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))).$$

Proof. 1. By Proposition 4.1.1, $\lambda(h, a).\nu_n(h) : H_n \times A \to \mathcal{P}(X)$ is a regular conditional distribution of $\mathbf{x}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. Then, by Proposition A.7.15, for all $n \in \mathbb{N}_0$, $\lambda(h, a).\nu_n(h) \odot \eta_{n+1} : H_n \times A \to \mathcal{P}(X)$ is a regular conditional distribution of $\mathbf{x}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

By Proposition 4.2.2, for all $n \in \mathbb{N}_0$, $\lambda(h, a, x).\mu_n(h, x) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$. Since τ_{n+1} is a regular conditional distribution, for all $n \in \mathbb{N}_0$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)) = \lambda\omega.\tau_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Proposition A.7.15 now shows that, for all $C \in \mathcal{Y}$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})) = \lambda\omega.(\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1})((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.}$$

That is, for all $n \in \mathbb{N}_0$, $\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1} : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

By Proposition A.7.12, $(\lambda(h, a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1}) : H_n \times A \to \mathcal{P}(X \times Y)$ is a regular conditional distribution of $(\mathbf{x}_{n+1}, \mathbf{y}_{n+1})$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. Then, by Proposition A.7.15, $((\lambda(h, a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1})) \odot \xi_{n+1} : H_n \times A \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. However, from Definition 2.2.2, $\Xi_{n+1} : H_n \times A \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. The result now follows from the uniqueness part of Proposition A.5.16.

2. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$,

$$\begin{aligned}
&\Xi_{n+1}(h_n, a_{n+1}) \\
&= (((\lambda(h,a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})) \odot \xi_{n+1})(h_n, a_{n+1}) \\
&= ((\lambda(h,a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1}))(h_n, a_{n+1}) \odot \\
&\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= ((\lambda(h,a).\nu_n(h) \odot \eta_{n+1})(h_n, a_{n+1}) \otimes \lambda x.(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})(h_n, a_{n+1}, x)) \odot \\
&\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= ((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})(h_n, a_{n+1}, x)) \odot \\
&\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= ((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y).
\end{aligned}$$

3. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$,

$$\begin{aligned}
&\breve{\Xi}_{n+1}(h_n, a_{n+1}) \cdot \upsilon_O \\
&= \Xi_{n+1}(h_n, a_{n+1}) \\
&= ((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= (((\breve{\nu}_n(h_n) \cdot \upsilon_X) \odot (\lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x) \cdot \upsilon_X)) \otimes \\
&\quad \lambda x.((\breve{\mu}_n(h_n, x) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y))) \odot \\
&\qquad (\lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \cdot \upsilon_X) \otimes \\
&\quad \lambda x.((\breve{\mu}_n(h_n, x) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y))) \odot \\
&\qquad (\lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&\qquad \text{[Proposition A.3.8]} \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \cdot \upsilon_X) \otimes \\
&\quad \lambda x.((\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y)) \odot \\
&\qquad (\lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&\qquad \text{[Proposition A.3.8]} \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \cdot \\
&\qquad (\upsilon_X \otimes \upsilon_Y)) \odot (\lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&\qquad \text{[Proposition A.12.3]} \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad \lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_O \\
&\qquad \text{[Proposition A.3.8]}
\end{aligned}$$

The result now follows by Proposition A.2.11.

4. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$, and all $C \in \mathcal{O}$,

$$\begin{aligned}
&\Xi_{n+1}(h_n, a_{n+1})(C) \\
&= (((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad\qquad \lambda(x, y).\xi_{n+1}(h_n, a_{n+1}, x, y))(C) \\
&= \int_{X \times Y} \lambda(x, y).\xi_{n+1}(h_n, a_{n+1}, x, y)(C) \\
&\qquad d((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))).
\end{aligned}$$

5. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$, and υ_O-almost all $o \in O$,

$$\begin{aligned}
&\breve{\Xi}_{n+1}(h_n, a_{n+1})(o) \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad\qquad \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y))(o) \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad\qquad ((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \\
&\qquad\qquad\qquad \lambda x.(\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \, d(\upsilon_X \otimes \upsilon_Y) \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad\qquad d\,((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \\
&\qquad\qquad\qquad \lambda x.(\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \cdot (\upsilon_X \otimes \upsilon_Y) \\
&\qquad\qquad\qquad\qquad \text{[Proposition A.3.3]} \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad\qquad d\,(((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \cdot \upsilon_X) \otimes \\
&\qquad\qquad\qquad (\lambda x.(\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y)) \\
&\qquad\qquad\qquad\qquad \text{[Proposition A.12.3]} \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad\qquad d\,(((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \cdot \upsilon_X) \otimes \\
&\qquad\qquad\qquad (\lambda x.((\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y))) \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad\qquad d\,(((\breve{\nu}_n(h_n) \cdot \upsilon_X) \odot (\lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x) \cdot \upsilon_X)) \otimes \\
&\qquad\qquad\qquad \lambda x.((\breve{\mu}_n(h_n, x) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y))) \\
&\qquad\qquad\qquad\qquad \text{[Proposition A.3.8]} \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad d((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))).
\end{aligned}$$

□

Part 1 of Proposition 4.2.4 shows that, for schemas $(\nu_n : H_n \to \mathcal{P}(X))_{n\in\mathbb{N}_0}$ and $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$, the environment can be synthesized from the schemas, their transition models, and the observation model for the conditional schema. Part 2 shows that, given the current history h_n and the action a_{n+1} just applied, the probability measure giving the distribution for the next observation o_{n+1} can be computed. Part 3 is similar, but for densities. Parts 4 and 5 give convenient ways of calculating $\Xi_{n+1}(h_n, a_{n+1})(C)$, for $C \in \mathcal{O}$, and $\breve{\Xi}_{n+1}(h_n, a_{n+1})(o)$, for $o \in O$.

By analogy with the case of states, suppose now that the conditional independence assumption

$$\sigma(\mathbf{o}_n) \perp\!\!\!\perp_{\sigma(\mathbf{x}_n,\mathbf{y}_n)} \sigma(\mathbf{h}_{n-1}, \mathbf{a}_n),$$

for all $n \in \mathbb{N}$, is satisfied. For all $n \in \mathbb{N}$, let $\xi'_n : X \times Y \to \mathcal{P}(O)$ be the regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{x}_n, \mathbf{y}_n)$. Hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{x}_{n+1}, \mathbf{y}_{n+1})) = \lambda\omega.\xi'_{n+1}((\mathbf{x}_{n+1}, \mathbf{y}_{n+1})(\omega))(B) \text{ a.s.}$$

Consider $\lambda(h, a, x, y).\xi'_{n+1}(x, y) : H_n \times A \times X \times Y \to \mathcal{P}(O)$. Now, for all $B \in \mathcal{O}$,

$$\lambda\omega.\xi'_{n+1}((\mathbf{x}_{n+1}, \mathbf{y}_{n+1})(\omega))(B) = \lambda\omega.(\lambda(h, a, x, y).\xi'_{n+1}(x, y))((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})(\omega))(B).$$

Since $\sigma(\mathbf{o}_{n+1}) \perp\!\!\!\perp_{\sigma(\mathbf{x}_{n+1},\mathbf{y}_{n+1})} \sigma(\mathbf{h}_n, \mathbf{a}_{n+1})$, it follows by Proposition A.6.1 that

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{x}_{n+1}, \mathbf{y}_{n+1})) = \mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})) \text{ a.s.},$$

for all $B \in \mathcal{O}$. Hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})) = \lambda\omega.(\lambda(h, a, x, y).\xi'_{n+1}(x, y))((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})(\omega))(B).$$

This means that $\lambda(h, a, x, y).\xi'_{n+1}(x, y)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{h}_n, \mathbf{y}_{n+1})$. It follows that

$$\xi_{n+1} = \lambda(h, a, x, y).\xi'_{n+1}(x, y) \ \ \mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1}))\text{-a.e.}$$

Thus, under the conditional independence assumption, the observation model can be simplified by dropping two arguments, the history and the action, from its domain. However, in the general case, the conditional independence assumption does not hold.

Thus it is necessary to handle functions like (a component of) a transition model, an observation model, or a schema that have the history space H_n as a factor of their domain. Of course, this is rather inconvenient, since for a start there are hardly any applications where it would be feasible to record the entire history throughout the lifetime of an agent. Instead what is needed is some kind of bounded summarization of the history. So let B_n denote the set which provides such summarizations of the histories in H_n and let $\beta_n : H_n \to B_n$ be the function which maps histories to their bounded summarization, for all $n \in \mathbb{N}_0$. Each element of B_n is called a *bounded history* up to time step n, for all

$n \in \mathbb{N}_0$. Now functions that have H_n as a factor of their domain need predicates defined over H_n for their definition. So what can be done is to choose predicates over H_n that can factored through B_n. Thus one considers predicates of the form $p \circ \beta_n : H_n \to \mathbb{B}$, where $p : B_n \to \mathbb{B}$ is a predicate over B_n. Predicates such as p are generated by a predicate grammar as discussed in Appendix B.1.

How can $\beta_n : H_n \to B_n$ be defined in such a way as to avoid having to store the history? One method is as follows. First define a function

$$\gamma_n : B_{n-1} \times A \times O \to B_n,$$

for all $n \in \mathbb{N}$. The function γ_n takes the current summarization and the next action a and observation o, and updates the summarization according to the action and observation. For example, if B_n is the window of the most recent 1000 action-observation pairs in the history, then (once the first 1000 actions in the lifetime of the agent have occurred) γ_n would throw away the oldest action-observation pair in the window and add (a, o) to the window. If B_n includes running statistics based on actions and observations that occur in the history, then γ_n would update the statistics according to the additional contributions from a and o.

Now, for all $n \in \mathbb{N}$, define

$$\beta_n : H_n \to B_n$$

by

$$\beta_n(h, a, o) = \gamma_n(\beta_{n-1}(h), a, o),$$

for all $h \in H_{n-1}$, $a \in A$, and $o \in O$. Also one needs to provide a suitable definition for $\beta_0 : H_0 \to B_0$. Then the β_n have the desired property.

Now a transition model can be defined so as to factor through the bounded history. Define

$$\overline{\tau}_n : B_{n-1} \times A \times X \times Y \to \mathcal{P}(Y)$$

to be a modified transition model that depends upon bounded history rather than history. Then define the transition model $\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y)$ by

$$\tau_n(h, a, x, y) = \overline{\tau}_n(\beta_{n-1}(h), a, x, y),$$

for all $h \in H_{n-1}$, $a \in A$, $x \in X$, and $y \in Y$.

Similarly, an observation model can be defined so as to factor through the bounded history. Define

$$\overline{\xi}_n : B_{n-1} \times A \times X \times Y \to \mathcal{P}(O)$$

to be a modified observation model that depends upon bounded history rather than history. Then define the observation model $\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O)$ by

$$\xi_n(h, a, x, y) = \overline{\xi}_n(\beta_{n-1}(h), a, x, y),$$

for all $h \in H_{n-1}$, $a \in A$, $x \in X$, and $y \in Y$.

Where do the transition and observation models come from? The most obvious answer is that they are given by the designer. For example, in robotics the situation is usually simple enough for the designer to know the definitions of the transition and observation models (except perhaps for learning some model parameters) and build these into the robot [156, Ch. 5 and 6]. In the general case, one could expect situations where it would be difficult for the designer to know exactly the definitions of the transition and observation models. In many cases, it would be reasonable for the designer to know the general form of these models but perhaps not all the details such as model parameters. Hence there is scope for using learning techniques to enable agents to learn accurate versions of the transition and observation models. Of course, the other situation where learning of the transition and observation models is needed is for applications where the transition and observation models naturally change over time due to changes in the environment.

Towards this, note that if an agent has accurate models, then Propositions 4.1.3 and 4.2.4 show that its predictions of what will happen next will turn out to be accurate because the environment can be synthesized using the transition and observation models. Put another way, if its predictions of what will happen next turn out to be inaccurate, the agent knows where the problem lies: there is a problem with its transition model, or its observation model, or both. And, for the setting of Proposition 4.2.4, the problem precisely is that either τ_n is not a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})$, or ξ_n is not a regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)$, or both, with an analogous remark for the setting of Proposition 4.1.3. At each time step, the agent can use the synthesized environment model given by Proposition 4.1.3 or 4.2.4, whichever is relevant, to calculate the likelihood of the observation. If the likelihood is 'high' (intuitively, the observation is near the mean), then nothing may need to be done; if the likelihood is 'low' (intuitively, the observation is in the tail of the distribution), then an adjustment may be needed.

Looking at the signatures of the transition and observation models, two places where adjustments from a learning process can take place are in the codomain $\mathcal{P}(Y)$ of the transition model and the codomain $\mathcal{P}(O)$ of the observation model. Now (each component of) both the transition model and the observation model is likely to be a piecewise-constant probability kernel. One idea is, at each time step, to adjust the parameters of the (finitely many) distributions that are values of the two probability kernels in such a way as to change the environment towards giving the observation actually observed a higher probability of occurring. An additional possiblity is to adjust the partitions to achieve the same effect.

The next result shows that, in an analogous way to the environment synthesis proposition (Proposition 4.2.4), it is possible to synthesize the observation model $(\zeta_n : H_{n-1} \times A \times X \to \mathcal{P}(O))_{n\in\mathbb{N}}$ for $\mathbf{x}$. The signature for the observation model is similar to that for the environment, but has an extra domain argument X. This means the equation defining the observation model turns out to be simpler than the one defining the environment.

Proposition 4.2.5. *(Observation model synthesis for the constant-valued, conditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes,*

$$(\zeta_n : H_{n-1} \times A \times X \to \mathcal{P}(O))_{n\in\mathbb{N}}$$

the observation model for $\mathbf{x}$*,*

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$$

the schema for $\mathbf{y}$ *given* $\mathbf{x}$*,*

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$$

the transition model for $\mathbf{y}$ *given* $\mathbf{x}$*,*

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$$

the observation model for $\mathbf{y}$ *given* $\mathbf{x}$*,* υ_O *a* σ*-finite measure on* $\mathcal{O}$*,* υ_X *a* σ*-finite measure on* $\mathcal{X}$*, and* υ_Y *a* σ*-finite measure on* $\mathcal{Y}$*. Suppose that, for all* $n \in \mathbb{N}_0$*, there is a conditional density* $\breve{\mu}_n : H_n \times X \to \mathcal{D}(Y)$ *such that* $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$*. Suppose that, for all* $n \in \mathbb{N}$*, there is a conditional density* $\breve{\zeta}_n : H_{n-1} \times A \times X \to \mathcal{D}(O)$ *such that* $\zeta_n = \breve{\zeta}_n \cdot \upsilon_O$*, a conditional density* $\breve{\tau}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(Y)$ *such that* $\tau_n = \breve{\tau}_n \cdot \upsilon_Y$*, and a conditional density* $\breve{\xi}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(O)$ *such that* $\xi_n = \breve{\xi}_n \cdot \upsilon_O$*. Suppose also that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n,\mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$*, and that* $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ *is constant-valued a.s. Then the following hold.*

1. *For all* $n \in \mathbb{N}_0$*,*

$$\zeta_{n+1} = (\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1}) \odot \xi_{n+1} \;\; \mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$*-almost all* $h_n \in H_n$*,* $a_{n+1} \in A$*, and* $x \in X$*,*

$$\zeta_{n+1}(h_n, a_{n+1}, x) = (\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y).$$

3. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$*-almost all* $h_n \in H_n$ $a_{n+1} \in A$*, and* $x \in X$*,*

$$\breve{\zeta}_{n+1}(h_n, a_{n+1}, x) = (\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \quad \upsilon_O\text{-a.e.}$$

4. *For all* $n \in \mathbb{N}_0$*,* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$*-almost all* $h_n \in H_n$*,* $a_{n+1} \in A$*, and* $x \in X$*, and all* $C \in \mathcal{O}$*,*

$$\zeta_{n+1}(h_n, a_{n+1}, x)(C) = \int_Y \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y)(C)\; d(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).$$

5. *For all* $n \in \mathbb{N}_0$*,* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$*-almost all* $h_n \in H_n$*,* $a_{n+1} \in A$*, and* $x \in X$*, and* υ_O*-almost all* $o \in O$*,*

$$\breve{\zeta}_{n+1}(h_n, a_{n+1}, x)(o) = \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\; d(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).$$

Proof. 1. By Proposition 4.2.2, for all $n \in \mathbb{N}_0$, $\lambda(h,a,x).\mu_n(h,x) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$. Since τ_{n+1} is a regular conditional distribution, for all $n \in \mathbb{N}_0$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)) = \lambda\omega.\tau_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Proposition A.7.15 now shows that, for all $C \in \mathcal{Y}$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})) = \lambda\omega.(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.}$$

That is, for all $n \in \mathbb{N}_0$, $\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1} : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Then, by Proposition A.7.15, $(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1}) \odot \xi_{n+1} : H_n \times A \times X \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$, for all $n \in \mathbb{N}_0$. However, from Definition 4.1.2, $\zeta_{n+1} : H_n \times A \times X \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$, for all $n \in \mathbb{N}_0$. The result now follows from the uniqueness part of Proposition A.5.16.

2. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1)})$-almost all $h_n \in H_n$, $a_{n+1} \in A$, and $x \in X$,

$$\begin{aligned}
&\zeta_{n+1}(h_n, a_{n+1}, x) \\
&= ((\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1}) \odot \xi_{n+1})(h_n, a_{n+1}, x) \\
&= (\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1})(h_n, a_{n+1}, x) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= (\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y).
\end{aligned}$$

3. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$-almost all $h_n \in H_n$, $a_{n+1} \in A$, and $x \in X$,

$$\begin{aligned}
&\breve{\zeta}_{n+1}(h_n, a_{n+1}, x) \cdot \upsilon_O \\
&= \zeta_{n+1}(h_n, a_{n+1}, x) \\
&= (\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= ((\breve{\mu}_n(h_n, x) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y)) \odot (\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&= ((\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&\qquad\qquad \text{[Proposition A.3.8]} \\
&= ((\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_O. \\
&\qquad\qquad \text{[Proposition A.3.8]}
\end{aligned}$$

The result now follows by Proposition A.2.11.

4. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$-almost all $h_n \in H_n$, $a_{n+1} \in A$, and $x \in X$, and all $C \in \mathcal{O}$,

$$\begin{aligned}
&\zeta_{n+1}(h_n, a_{n+1}, x)(C) \\
&= ((\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y))(C) \\
&= \int_Y \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y)(C)\ d(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).
\end{aligned}$$

5. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$-almost all $h_n \in H_n$, $a_{n+1} \in A$, and $x \in X$, and υ_O-almost all $o \in O$,

$$\begin{aligned}
&\breve{\zeta}_{n+1}(h_n, a_{n+1}, x)(o) \\
&= ((\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y))(o) \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ (\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))\ d\upsilon_Y \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ \ d((\breve{\mu}_n(h_n, x) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[Proposition A.3.3]} \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ \ d((\breve{\mu}_n(h_n, x) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y)) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[Proposition A.3.8]} \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ d(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).
\end{aligned}$$

□

Part 1 of Proposition 4.2.5 shows that the observation model $(\zeta_n : H_{n-1} \times A \times X \to \mathcal{P}(O))_{n\in\mathbb{N}}$ can be synthesized from the schema $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$, and its transition and observation models. Part 5 gives a convenient way of calculating $\breve{\zeta}_{n+1}(h_n, a_{n+1}, x)(o)$, for $o \in O$. This will crucial for designing particle filters that learn unknown parameters in Section 4.4.

The equations

$$\begin{aligned}
\Xi_{n+1} &= (\lambda(h, a).\nu_n(h) \odot \eta_{n+1}) \odot \zeta_{n+1} \\
\Xi_{n+1} &= ((\lambda(h, a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1})) \odot \xi_{n+1} \\
\zeta_{n+1} &= (\lambda(h, a, x).\mu_n(h, x) \odot \tau_{n+1}) \odot \xi_{n+1}
\end{aligned}$$

of Proposition 4.1.3, Proposition 4.2.4, and Proposition 4.2.5, respectively, are intrinsic properties of agent-environment systems. Other such properties are given by Proposition 4.1.4 above and Propositions 4.2.6 below. They come about because of the requirement that environments, schemas, and transition and observation models must all be regular conditional distributions which restricts their possible definitions. The above properties are useful in practice; recall the comments about learning transition and observation models after Proposition 4.2.4 and see the application of Proposition 4.2.5 to filtering in Section 4.2.

Example 4.2.1. Consider filtering a schema of the form $(\mu_n : H_n \to \mathcal{P}(X_1 \times X_2 \times X_3))_{n\in\mathbb{N}}$. Here are the signatures for the components of the schema, transition model, and observation model.

$$\begin{aligned}
&\mu_n : H_n \to \mathcal{P}(X_1 \times X_2 \times X_3) \\
&\tau_n : H_{n-1} \times A \times X_1 \times X_2 \times X_3 \to \mathcal{P}(X_1 \times X_2 \times X_3) \\
&\xi_n : H_{n-1} \times A \times X_1 \times X_2 \times X_3 \to \mathcal{P}(O)
\end{aligned}$$

One can filter directly with these models using Proposition 4.1.2.

Alternatively, μ can be deconstructed into three factor schemas whose components have signatures $\mu_n^{(1)} : H_n \to \mathcal{P}(X_1)$, $\mu_n^{(2)} : H_n \times X_1 \to \mathcal{P}(X_2)$, and $\mu_n^{(3)} : H_n \times X_1 \times X_2 \to \mathcal{P}(X_3)$. Then $\mu_n^{(1)}$, $\mu_n^{(2)}$, and $\mu_n^{(3)}$ can be filtered separately. For each of these schemas, here are the signatures for the schema, transition model, and observation model.

$$\begin{aligned}
&\mu_n^{(1)} : H_n \to \mathcal{P}(X_1)\\
&\tau_n^{(1)} : H_{n-1} \times A \times X_1 \to \mathcal{P}(X_1)\\
&\xi_n^{(1)} : H_{n-1} \times A \times X_1 \to \mathcal{P}(O)
\end{aligned}$$

$$\begin{aligned}
&\mu_n^{(2)} : H_n \times X_1 \to \mathcal{P}(X_2)\\
&\tau_n^{(2)} : H_{n-1} \times A \times X_1 \times X_2 \to \mathcal{P}(X_2)\\
&\xi_n^{(2)} : H_{n-1} \times A \times X_1 \times X_2 \to \mathcal{P}(O)
\end{aligned}$$

$$\begin{aligned}
&\mu_n^{(3)} : H_n \times X_1 \times X_2 \to \mathcal{P}(X_3)\\
&\tau_n^{(3)} : H_{n-1} \times A \times X_1 \times X_2 \times X_3 \to \mathcal{P}(X_3)\\
&\xi_n^{(3)} : H_{n-1} \times A \times X_1 \times X_2 \times X_3 \to \mathcal{P}(O).
\end{aligned}$$

Now, using Proposition 4.2.5 for the observation model $\xi^{(1)}$ and a similar argument for the observation model $\xi^{(2)}$, each can be synthesized as follows:

$$\begin{aligned}
\xi_{n+1}^{(1)} &= (\mu_n^{(2)^\dagger} \odot \tau_{n+1}^{(2)}) \odot \xi_{n+1}^{(2)}\\
\xi_{n+1}^{(2)} &= (\mu_n^{(3)^\dagger} \odot \tau_{n+1}^{(3)}) \odot \xi_{n+1}^{(3)}.
\end{aligned}$$

Thus, of the three observation models, only the observation model $\xi^{(3)}$ needs to be known or learned since the other two can be calculated from $\xi^{(3)}$ (and some other ingredients).

For simulation, the transition and observation models relevant to μ should be used. Thus, for the observation model, ξ, which is the same as $\xi^{(3)}$, should be used. The other two observation models, $\xi^{(1)}$ and $\xi^{(2)}$, are not needed for simulation. Obviously these remarks extend to any product $\prod_{i=1}^m X_i$: only $\xi^{(m)}$ needs to be known or learned for filtering and only $\xi^{(m)}$ is needed for simulation.

Some actions can be no-ops.

Example 4.2.2. Let $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ be a schema. Consider a transition model $(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$ and $a \in A$ such that $\tau_n(h_{n-1}, a, x, y) = \delta_y$, for all $h_{n-1} \in H_{n-1}$, $x \in X$, $y \in Y$, and $n \in \mathbb{N}$. By Proposition A.2.7,

$$\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a, x, y) = \mu_n(h_n, x),$$

for all $h_n \in H_n$, $x \in X$, and $n \in \mathbb{N}$. In other words, for an action having such a transition model, there is no change to the empirical belief during the transition update. That is, the action is a *no-op*.

An observation model may be non-informative.

Example 4.2.3. Suppose that the observation model $(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$ satisfies the condition that, for all $h_n \in H_n$, $a_{n+1} \in A$, $x \in X$, $o_{n+1} \in O$, and $n \in \mathbb{N}$, $\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) : Y \to \mathbb{R}$ is a constant function, which may vary with h_n, a_{n+1}, x, o_{n+1} and n. This is the case, for example, if O is a singleton set. Then

$$\begin{aligned}
&\lambda x.\mu_{n+1}(h_{n+1}, x) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) \ * \ \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \\
&= \lambda x.(\mu_n(h_n, x) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).
\end{aligned}$$

Thus ξ is an observation model that provides no information for the observation update, which therefore does not change the empirical belief produced by the transition update. Such an observation model is called *non-informative*.

Example 4.2.4. Consider an application where it is required to acquire an empirical belief in $\mathcal{P}(G)$, where the empirical belief is a distribution over a set G of graphs, say, and depends on some (hidden) parameters that live in a parameter space P, say. This example shows how the parameters may be acquired and then marginalized out to obtain the desired empirical belief.

Consider the schema $(\mu_n)_{n\in\mathbb{N}_0}$, where

$$\mu_n : H_n \to \mathcal{P}(P \times G).$$

This product schema deconstructs into two schemas with respective components

$$\begin{aligned}
&\mu_n^{(1)} : H_n \to \mathcal{P}(P) \\
&\mu_2^{(2)} : H_n \times P \to \mathcal{P}(G),
\end{aligned}$$

so that $\mu_n = \mu_n^{(1)} \otimes \mu_n^{(2)}$, for all $n \in \mathbb{N}_0$.

Consider the filtering problem for $(\mu_n^{(1)} : H_n \to \mathcal{P}(P))_{n\in\mathbb{N}_0}$. The components for the transition model have signature

$$\tau_n^{(1)} : H_{n-1} \times A \times P \to \mathcal{P}(P),$$

while the components for the observation model have signature

$$\xi_n^{(1)} : H_{n-1} \times A \times P \to \mathcal{P}(O).$$

Consider now the filtering problem for $(\mu_n^{(2)} : H_n \times P \to \mathcal{P}(G))_{n\in\mathbb{N}_0}$. The components for the transition model have signature

$$\tau_n^{(2)} : H_{n-1} \times A \times P \times G \to \mathcal{P}(G),$$

while the components for the observation model have signature

$$\xi_n^{(2)} : H_{n-1} \times A \times P \times G \to \mathcal{P}(O).$$

Then, at any time, one can use $\mu_n^{(1)}(h_n) : \mathcal{P}(P)$ and $\lambda p.\mu_n^{(2)}(h_n, p) : P \to \mathcal{P}(G)$ to construct $\mu_n(h_n) : \mathcal{P}(P \times G)$ via $\mu_n(h_n) = \mu_n^{(1)}(h_n) \otimes \lambda p.\mu_n^{(2)}(h_n, p)$. However, for choosing actions, it is more likely that the marginal distribution over the graphs *alone* will be needed. This can be calculated using fusion of probability kernels via

$$\mu_n^{(1)}(h_n) \odot \lambda p.\mu_n^{(2)}(h_n, p) : \mathcal{P}(G).$$

To see this, let $(\Omega, \mathfrak{S}, \mathsf{P})$ be the basic probability space, $(A, \mathcal{A})$ the action space, $(O, \mathcal{O})$ the observation space, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ the action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ the observation process, and $\mathbf{p} : \Omega \to P^{\mathbb{N}_0}$ and $\mathbf{g} : \Omega \to G^{\mathbb{N}_0}$ the relevant stochastic processes. (See Figure 4.13.) Then Proposition A.7.15 shows that $\mu_n^{(1)} \odot \mu_n^{(2)} : H_n \to \mathcal{P}(G)$ is a regular conditional distribution of $\mathbf{g}_n$ given $\mathbf{h}_n$, for all $n \in \mathbb{N}_0$. It follows that $(\mu_n^{(1)} \odot \mu_n^{(2)})_{n \in \mathbb{N}_0}$ is the schema for $\mathbf{g}$ given $\mathbf{h}$ and so $\mu_n^{(1)}(h_n) \odot \lambda p.\mu_n^{(2)}(h_n, p)$ is the desired empirical belief.

In addition, recall that

$$\mu_n^{(1)}(h_n) \odot \lambda p.\mu_n^{(2)}(h_n, p) = (\mu_n^{(1)}(h_n) \otimes \lambda p.\mu_n^{(2)}(h_n, p)) \circ \pi_G^{-1},$$

where $\pi_G : P \times G \to G$ is the canonical projection. Hence $\mu_n^{(1)}(h_n) \odot \lambda p.\mu_n^{(2)}(h_n, p)$ is the marginal probability measure for $\mu_n^{(1)}(h_n) \otimes \lambda p.\mu_n^{(2)}(h_n, p)$ with respect to G, by Definitions A.2.7 and A.2.9, and the note immediately after Proposition A.7.1.

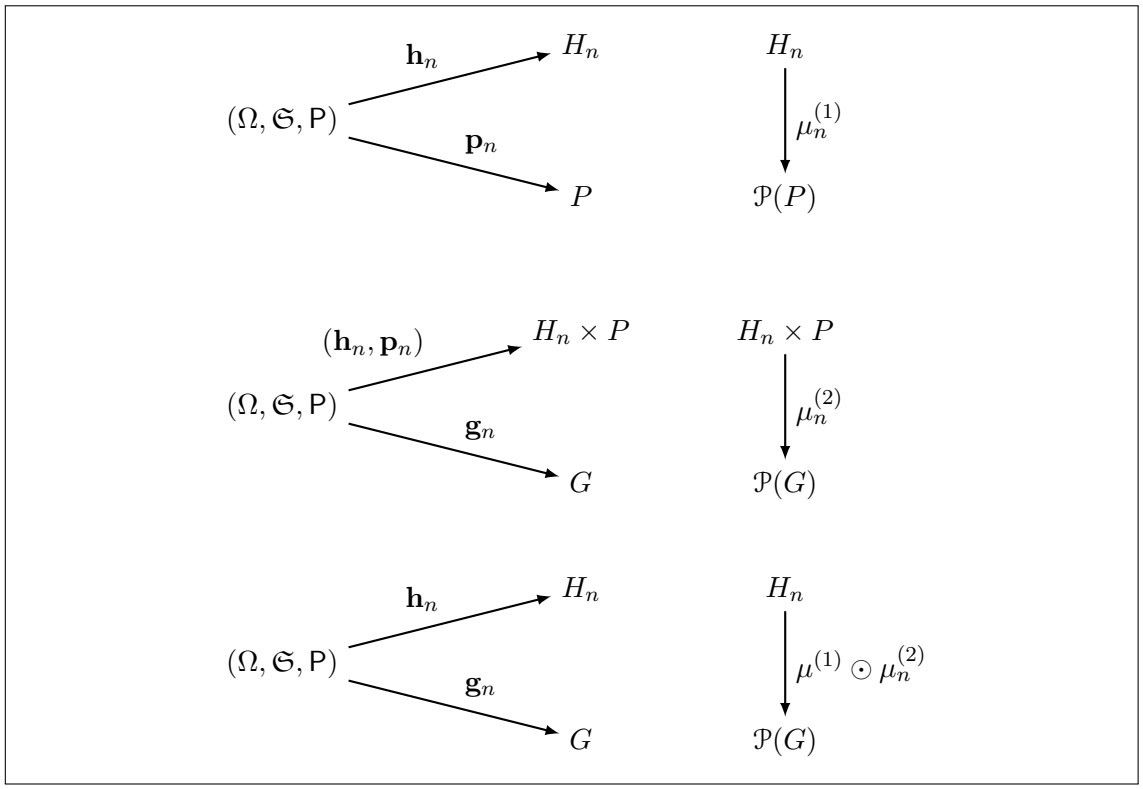

Figure 4.13: Setting for Example 4.2.4

When part of the space supporting the probability measures in the codomain of an empirical belief is a bunch of parameters like this, it is always possible to marginalize out the parameters to obtain the empirical belief relevant to the rest of the space.

Of course, it is also possible to try to directly acquire empirical beliefs about G without use of the parameters in P. Towards this, consider a schema $(\mu_n^G : H_n \to \mathcal{P}(G))_{n\in\mathbb{N}_0}$ for $\mathbf{g}$ given $\mathbf{h}$, and its filtering problem. The components for the transition model have signature

$$\tau_n^{(G)} : H_{n-1} \times A \times G \to \mathcal{P}(G),$$

while the components for the observation model have signature

$$\xi_n^{(G)} : H_{n-1} \times A \times G \to \mathcal{P}(O).$$

Furthermore, by the essential uniqueness of schemas, $\mu_n^{(G)} = \mu_n^{(1)} \odot \mu^{(2)}$ almost everywhere, for all $n \in \mathbb{N}_0$. Whether or not it is better to use the direct approach or else to introduce some (hidden) parameters and then marginalize them out depends upon the application.

Now comes Bayes theorem for schemas. This result extends Proposition 4.1.4 to the conditional case for which $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. and is a special case of Proposition A.7.14.

Proposition 4.2.6. *(Bayes theorem for conditional schemas) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes, $(\zeta_n : H_{n-1} \times A \times X \to \mathcal{P}(O))_{n\in\mathbb{N}_0}$ the observation model for $\mathbf{x}$, $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ the schema for $\mathbf{y}$ given $\mathbf{x}$, $(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$ the transition model for $\mathbf{y}$ given $\mathbf{x}$, and $(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$ the observation model for $\mathbf{y}$ given $\mathbf{x}$. Suppose that $\sigma(\mathbf{a}_{n+1}) \perp\!\!\!\perp_{\sigma(\mathbf{h}_n,\mathbf{x}_n)} \sigma(\mathbf{y}_n)$, for all $n \in \mathbb{N}_0$, and that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. Then, for all $n \in \mathbb{N}_0$,*

$$\lambda(h,a,x).(\zeta_{n+1} \otimes \lambda(h,a,x,o).\mu_{n+1}(h,a,o,x))(h,a,x)(E^*) =$$
$$\lambda(h,a,x).((\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1}) \otimes \xi_{n+1})(h,a,x)(E) \;\; \mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))\text{-a.e.},$$

for all $E \in \mathcal{Y} \otimes \mathcal{O}$.

Proof. As shown in the proof of Proposition 4.2.3, for all $n \in \mathbb{N}_0$, $\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1} : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$. Also $\xi_{n+1} : H_n \times A \times X \times Y \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})$. Hence, by Proposition A.7.12,

$$(\lambda(h,a,x).\mu_n(h,x) \odot \tau_{n+1}) \otimes \xi_{n+1} : H_n \times A \times X \to \mathcal{P}(Y \times O)$$

is a regular conditional distribution of $(\mathbf{y}_{n+1}, \mathbf{o}_{n+1})$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Now, for all $n \in \mathbb{N}$, $\zeta_{n+1} : H_n \times A \times X \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$. Also $\mu_{n+1} : H_n \times A \times O \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{o}_{n+1}, \mathbf{x}_{n+1})$. Thus $\lambda(h,a,x,o).\mu_{n+1}(h,a,o,x) : H_n \times A \times X \times O \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{o}_{n+1})$. Hence, by Proposition A.7.12,

$$\zeta_{n+1} \otimes \lambda(h,a,x,o).\mu_{n+1}(h,a,o,x) : H_n \times A \times X \to \mathcal{P}(O \times Y)$$

is a regular conditional distribution of $(\mathbf{o}_{n+1}, \mathbf{y}_{n+1})$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

The result now follows by the uniqueness part of Proposition A.5.16. □

The following result for this section provides a theoretical foundation for simulation in the conditional case.

Proposition 4.2.7. *(Simulation for the conditional case) Let $(A, \mathcal{A})$ be an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces,*

$$\nu_0 : \mathcal{P}(X)$$

a probability measure,

$$\mu_0 : X \to \mathcal{P}(Y)$$

a probability kernel,

$$(\Lambda_n : H_{n-1} \to \mathcal{P}(A))_{n\in\mathbb{N}},$$
$$(\eta_n : H_{n-1} \times A \times X \to \mathcal{P}(X))_{n\in\mathbb{N}},$$
$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}},$$

and

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$$

sequences of probability kernels. Then there exists a probability space $(\Omega, \mathfrak{S}, \mathsf{P})$, an action process $\mathbf{a} : \Omega \to A^{\mathbb{N}}$, an observation process $\mathbf{o} : \Omega \to O^{\mathbb{N}}$, and stochastic processes $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ such that

$$\sigma(\mathbf{a}_n) \underset{\sigma(\mathbf{h}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}),$$
$$\sigma(\mathbf{x}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{y}_1, \mathbf{x}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n),$$
$$\sigma(\mathbf{y}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{x}_n),$$

and

$$\sigma(\mathbf{o}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n),$$

for all $n \in \mathbb{N}$. Furthermore,

$$(\Lambda_n : H_{n-1} \to \mathcal{P}(A))_{n\in\mathbb{N}}$$

is the agent for $\mathbf{a}$ and $\mathbf{o}$,

$$(\eta_n : H_{n-1} \times A \times X \to \mathcal{P}(X))_{n\in\mathbb{N}}$$

the transition model for $\mathbf{x}$,

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}}$$

the transition model for $\mathbf{y}$ given $\mathbf{x}$, and

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$$

the observation model for $\mathbf{y}$ given $\mathbf{x}$.

Proof. Let $\Omega \triangleq X \times Y \times A \times X \times Y \times O \times A \times X \times Y \times O \times \cdots$ and let $\mathfrak{S}$ be the usual product σ-algebra on Ω. Also let $\tilde{H}_n \triangleq X \times Y \times (A \times X \times Y \times O)^n$, for all $n \in \mathbb{N}_0$, and give each $\tilde{H}_n$ the usual product σ-algebra.

Define, for all $n \in \mathbb{N}$,

$$\begin{aligned}
\tilde{\Lambda}_n &\triangleq \lambda(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_{n-1}, x_{n-1}, y_{n-1}, o_{n-1}).\Lambda_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}) \\
&\qquad : \tilde{H}_{n-1} \to \mathcal{P}(A) \\
\tilde{\eta}_n &\triangleq \lambda(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_{n-1}, x_{n-1}, y_{n-1}, o_{n-1}, a_n). \\
&\qquad \eta_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}, a_n, x_{n-1}) : \tilde{H}_{n-1} \times A \to \mathcal{P}(X) \\
\tilde{\tau}_n &\triangleq \lambda(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_{n-1}, x_{n-1}, y_{n-1}, o_{n-1}, a_n, x_n). \\
&\qquad \tau_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}, a_n, x_n, y_{n-1}) : \tilde{H}_{n-1} \times A \times X \to \mathcal{P}(Y) \\
\tilde{\xi}_n &\triangleq \lambda(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_{n-1}, x_{n-1}, y_{n-1}, o_{n-1}, a_n, x_n, y_n). \\
&\qquad \xi_n(a_1, o_1, \ldots, a_{n-1}, o_{n-1}, a_n, x_n, y_n) : \tilde{H}_{n-1} \times A \times X \times Y \to \mathcal{P}(O).
\end{aligned}$$

Then each $\tilde{\Lambda}_n$, $\tilde{\eta}_n$, $\tilde{\tau}_n$, and $\tilde{\xi}_n$ is a probability kernel.

Define $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ to be the canonical projection. Hence $\mathbf{a}$ is an action process based on A. Also, define $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ to be the canonical projections. Hence $\mathbf{x}$ and $\mathbf{y}$ are stochastic processes. Similarly, define $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ to be the canonical projection. Hence $\mathbf{o}$ is an observation process based on O.

By Proposition A.8.1, there exists a unique probability measure P on $(\Omega, \mathfrak{S})$ such that

$$\begin{aligned}
&\mathsf{P} \circ \mathbf{x}_0^{-1} = \nu_0 \\
&\mathsf{P} \circ (\mathbf{x}_0, \mathbf{y}_0)^{-1} = \nu_0 \otimes \mu_0
\end{aligned}$$

and, for all $n \in \mathbb{N}$,

$$\begin{aligned}
&\mathsf{P} \circ (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n)^{-1} = \nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \\
&\mathsf{P} \circ (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{x}_n)^{-1} = \nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \otimes \tilde{\eta}_n \\
&\mathsf{P} \circ (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)^{-1} = \\
&\qquad \nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \otimes \tilde{\eta}_n \otimes \tilde{\tau}_n \\
&\mathsf{P} \circ (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n, \mathbf{o}_n)^{-1} = \\
&\qquad \nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \otimes \tilde{\eta}_n \otimes \tilde{\tau}_n \otimes \tilde{\xi}_n.
\end{aligned}$$

In particular, $(\Omega, \mathfrak{S}, \mathsf{P})$ is a probability space.

Now it is shown that, for all $n \in \mathbb{N}$,

$$\sigma(\mathbf{a}_n) \underset{\sigma(\mathbf{h}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}).$$

Towards this, for all $n \in \mathbb{N}$ and $C \in \mathcal{A}$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}(\mathbf{a}_n^{-1}(C) \mid (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1})) \\
&= \lambda\omega.\tilde{\Lambda}_n((\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1})(\omega))(C) \\
&\qquad\qquad \text{[Proposition A.8.2]} \\
&= \lambda\omega.\Lambda_n(\mathbf{h}_{n-1}(\omega))(C) \\
&= \mathsf{P}(\mathbf{a}_n^{-1}(C) \mid \mathbf{h}_{n-1}). \qquad \text{[Proposition A.7.18]}
\end{aligned}$$

Hence the result. Furthermore, the first step above shows that each $\tilde{\Lambda}_n$ is a regular conditional distribution and the last step shows that Λ is the agent for $\mathbf{a}$ and $\mathbf{o}$.

It is shown that, for all $n \in \mathbb{N}$,

$$\sigma(\mathbf{x}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n).$$

Towards this, for all $n \in \mathbb{N}$ and $D \in \mathcal{X}$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}(\mathbf{x}_n^{-1}(D) \mid (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n)) \\
&= \lambda\omega.\tilde{\eta}_n((\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n)(\omega))(D) \\
&\qquad\qquad \text{[Proposition A.8.2]} \\
&= \lambda\omega.\eta_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_{n-1})(\omega))(D) \\
&= \mathsf{P}(\mathbf{x}_n^{-1}(D) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_{n-1})). \qquad \text{[Proposition A.7.18]}
\end{aligned}$$

Hence the result. Furthermore, the first step above shows that each $\tilde{\eta}_n$ is a regular conditional distribution and the last step shows that η is the transition model for $\mathbf{x}$.

Next it is shown that, for all $n \in \mathbb{N}$,

$$\sigma(\mathbf{y}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})}{\perp\!\!\!\perp} \sigma(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{x}_n).$$

Towards this, for all $n \in \mathbb{N}$ and $E \in \mathcal{Y}$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}(\mathbf{y}_n^{-1}(E) \mid (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{x}_n)) \\
&= \lambda\omega.\tilde{\tau}_n((\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{x}_n)(\omega))(E) \\
&\qquad\qquad \text{[Proposition A.8.2]} \\
&= \lambda\omega.\tau_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})(\omega))(E) \\
&= \mathsf{P}(\mathbf{y}_n^{-1}(E) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1})). \qquad \text{[Proposition A.7.18]}
\end{aligned}$$

Hence the result. Furthermore, the first step above shows that each $\tilde{\tau}_n$ is a regular conditional distribution and the last step shows that τ is the transition model for $\mathbf{y}$ given $\mathbf{x}$.

Finally, it is shown that, for all $n \in \mathbb{N}$,

$$\sigma(\mathbf{o}_n) \underset{\sigma(\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n).$$

Towards this, for all $n \in \mathbb{N}$ and $F \in \mathcal{O}$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}(\mathbf{o}_n^{-1}(F) \mid (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)) \\
&= \lambda\omega.\tilde{\xi}_n((\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_{n-1}, \mathbf{x}_{n-1}, \mathbf{y}_{n-1}, \mathbf{o}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)(\omega))(F) \\
&\qquad\qquad \text{[Proposition A.8.2]} \\
&= \lambda\omega.\xi_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n\mathbf{y}_n)(\omega))(F) \\
&= \mathsf{P}(\mathbf{o}_n^{-1}(F) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n)). \qquad \text{[Proposition A.7.18]}
\end{aligned}$$

Hence the result. Furthermore, the first step above shows that each $\tilde{\xi}_n$ is a regular conditional distribution and the last step shows that ξ is the observation model for $\mathbf{y}$ given $\mathbf{x}$. □

The definitions of ν_0, μ_0, Λ, τ, and ξ in Proposition 4.2.7 can be completely arbitrary. The result thus provide the basis for a simulation of an agent-environment system for which the choices of initial empirical beliefs, agent, transition models, and observation model are arbitrary.

Finally, in this subsection, it is shown that if the transition model η in a simulation is such that every action is a no-op, then $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s.

Proposition 4.2.8. *Under the conditions of Proposition 4.2.7, suppose in addition that $\eta_n = \lambda(h, a, x).\delta_x$, for all $n \in \mathbb{N}$. Then $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s.*

Proof. To begin with, it is shown that

$$\nu_0(\{x_0 \in X\}) = 1$$
$$(\nu_0 \otimes \mu_0)(\{(x_0, y_0) \in X \times Y\}) = 1$$

and, for all $n \in \mathbb{N}$,

$$\begin{aligned}
&(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n) \\
&\quad (\{(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_n) \\
&\qquad \in X \times Y \times (A \times X \times Y \times O)^{n-1} \times A \mid x_0 = x_1 = \cdots = x_{n-1}\}) = 1 \\
&(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \otimes \tilde{\eta}_n) \\
&\quad (\{(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_n, x_n) \\
&\qquad \in X \times Y \times (A \times X \times Y \times O)^{n-1} \times A \times X \mid x_0 = x_1 = \cdots = x_n\}) = 1 \\
&(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \otimes \tilde{\eta}_n \otimes \tilde{\tau}_n) \\
&\quad (\{(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_n, x_n, y_n) \\
&\qquad \in X \times Y \times (A \times X \times Y \times O)^{n-1} \times A \times X \times Y \mid x_0 = x_1 = \cdots = x_n\}) = 1 \\
&(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1 \otimes \tilde{\tau}_1 \otimes \tilde{\xi}_1 \otimes \cdots \otimes \tilde{\Lambda}_n \otimes \tilde{\eta}_n \otimes \tilde{\tau}_n \otimes \tilde{\xi}_n) \\
&\quad (\{(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_n, x_n, y_n, o_n) \\
&\qquad \in X \times Y \times (A \times X \times Y \times O)^{n} \mid x_0 = x_1 = \cdots = x_n\}) = 1.
\end{aligned}$$

The first two parts are obvious, since $\nu_0(X) = 1$ and $(\nu_0 \otimes \mu_0)(X \times Y) = 1$.

The remaining parts are proved by induction. Consider first the base case, when $n = 1$. Clearly, $(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1)(\{(x_0, y_0, a_1) \in X \times Y \times A\}) = 1$.

Next

$$\begin{aligned}
&(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1 \otimes \tilde{\eta}_1)(\{(x_0, y_0, a_1, x_1) \in X \times Y \times A \times X \mid x_0 = x_1\}) \\
&= \int_{X \times Y \times A} (\lambda(x_0, y_0, a_1). \int_X \lambda x_1.\mathbf{1}_{\{(x_0,y_0,a_1,x_1) \in X \times Y \times A \times X \mid x_0 = x_1\}}(x_0, y_0, a_1, x_1) \\
&\qquad\qquad d\tilde{\eta}_1(x_0, y_0, a_1))\ d(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1) \\
&= \int_{X \times Y \times A} (\lambda(x_0, y_0, a_1). \int_X \lambda x_1.\mathbf{1}_{\{(x_0,y_0,a_1,x_1) \in X \times Y \times A \times X \mid x_0 = x_1\}}(x_0, y_0, a_1, x_1)\ d\delta_{x_0}) \\
&\qquad\qquad d(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1) \\
&= \int_{X \times Y \times A} \lambda(x_0, y_0, a_1).\mathbf{1}_{\{(x_0,y_0,a_1,x_1) \in X \times Y \times A \times X \mid x_0 = x_1\}}(x_0, y_0, a_1, x_0)\ d(\nu_0 \otimes \mu_0 \otimes \tilde{\Lambda}_1)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{X\times Y\times A} \mathbf{1}_{\{(x_0,y_0,a_1)\in X\times Y\times A\}}\, d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1)\\
&= (\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1)(\{(x_0,y_0,a_1)\in X\times Y\times A\})\\
&= 1. && \text{[Preceding part]}
\end{aligned}
$$

Also

$$
\begin{aligned}
&(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1)(\{(x_0,y_0,a_1,x_1,y_1)\in X\times Y\times A\times X\times Y \mid x_0=x_1\})\\
&= (\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1)(\{(x_0,y_0,a_1,x_1)\in X\times Y\times A\times X \mid x_0=x_1\}\times Y)\\
&= \int_{X\times Y\times A\times X} \mathbf{1}_{\{(x_0,y_0,a_1,x_1)\in X\times Y\times A\times X \mid x_0=x_1\}}\, \lambda(x_0,y_0,a_1,x_1).\tilde{\tau}_1(x_0,y_0,a_1,x_1)(Y)\\
&\qquad\qquad d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1)\\
&= \int_{X\times Y\times A\times X} \mathbf{1}_{\{(x_0,y_0,a_1,x_1)\in X\times Y\times A\times X \mid x_0=x_1\}}\, d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1)\\
&= (\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1)(\{(x_0,y_0,a_1,x_1)\in X\times Y\times A\times X \mid x_0=x_1\})\\
&= 1. && \text{[Preceding part]}
\end{aligned}
$$

The proof of the final part for the base case is similar to the preceding one. This completes the base case.

For the induction step, suppose that each of the parts hold for $n-1$. Then

$$
\begin{aligned}
&(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\Lambda}_n)\\
&\qquad\qquad (\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n) \mid x_0=x_1=\cdots=x_{n-1}\})\\
&= (\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\Lambda}_n)\\
&\qquad\qquad (\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1}) \mid x_0=x_1=\cdots=x_{n-1}\}\times A)\\
&= \int_{X\times Y\times(A\times X\times Y\times O)^{n-1}} \mathbf{1}_{\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1}) \mid x_0=x_1=\cdots=x_{n-1}\}}\\
&\qquad \lambda(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1}).\tilde{\Lambda}_n(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1})(A)\\
&\qquad\qquad d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\xi}_{n-1})\\
&= \int_{X\times Y\times(A\times X\times Y\times O)^{n-1}} \mathbf{1}_{\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1}) \mid x_0=x_1=\cdots=x_{n-1}\}}\\
&\qquad\qquad d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\xi}_{n-1})\\
&= (\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\xi}_{n-1})\\
&\qquad\qquad (\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1}) \mid x_0=x_1=\cdots=x_{n-1}\})\\
&= 1. && \text{[Induction hypothesis]}
\end{aligned}
$$

Next

$$
\begin{aligned}
&(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\Lambda}_n\otimes\tilde{\eta}_n)\\
&\qquad\qquad (\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n,x_n) \mid x_0=x_1=\cdots=x_n\})
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{X\times Y\times(A\times X\times Y\times O)^{n-1}\times A} (\lambda(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n).\\
&\qquad \int_X \lambda x_n.\mathbf{1}_{\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n,x_n)\mid x_0=x_1=\cdots=x_n\}}(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n,x_n)\\
&\qquad\qquad d\tilde{\eta}_n(x_0,\ldots,o_{n-1},a_n))\; d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\Lambda}_n)\\
&= \int_{X\times Y\times(A\times X\times Y\times O)^{n-1}\times A} (\lambda(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n).\\
&\qquad \int_X \lambda x_n.\mathbf{1}_{\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n,x_n)\mid x_0=x_1=\cdots=x_n\}}(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n,x_n)\\
&\qquad\qquad d\delta_{x_{n-1}})\; d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\Lambda}_n)\\
&= \int_{X\times Y\times(A\times X\times Y\times O)^{n-1}\times A} \lambda(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n).\\
&\quad \mathbf{1}_{\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n,x_n)\mid x_0=x_1=\cdots=x_n\}}(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n,x_{n-1})\\
&\qquad\qquad d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\Lambda}_n)\\
&= \int_{X\times Y\times(A\times X\times Y\times O)^{n-1}\times A} \lambda(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n).\\
&\quad \mathbf{1}_{\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n)\mid x_0=x_1=\cdots=x_{n-1}\}}(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n)\\
&\qquad\qquad d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\tilde{\Lambda}_n)\\
&= \int_{X\times Y\times(A\times X\times Y\times O)^{n-1}\times A} \mathbf{1}_{\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n)\mid x_0=x_1=\cdots=x_{n-1}\}}\\
&\qquad\qquad d(\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\Lambda}_n)\\
&= (\nu_0\otimes\mu_0\otimes\tilde{\Lambda}_1\otimes\tilde{\eta}_1\otimes\tilde{\tau}_1\otimes\tilde{\xi}_1\otimes\cdots\otimes\tilde{\Lambda}_n)\\
&\qquad\qquad (\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,o_{n-1},a_n)\mid x_0=x_1=\cdots=x_{n-1}\})\\
&= 1. \qquad\qquad \text{[Preceding part]}
\end{aligned}
$$

The remaining two cases are similar to the first one. This completes the induction proof.

Proposition A.8.1 now shows that

$$
\begin{aligned}
&(\mathsf{P}\circ\mathbf{x}_0^{-1})(\{x_0\in X\})=1\\
&(\mathsf{P}\circ(\mathbf{x}_0,\mathbf{y}_0)^{-1})(\{(x_0,y_0)\in X\times Y\})=1
\end{aligned}
$$

and, for all $n\in\mathbb{N}$,

$$
\begin{aligned}
&(\mathsf{P}\circ(\mathbf{x}_0,\mathbf{y}_0,\mathbf{a}_1,\mathbf{x}_1,\mathbf{y}_1,\mathbf{o}_1,\ldots,\mathbf{a}_n)^{-1})\\
&\qquad\qquad (\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n)\mid x_0=x_1=\cdots=x_{n-1}\})=1\\
&(\mathsf{P}\circ(\mathbf{x}_0,\mathbf{y}_0,\mathbf{a}_1,\mathbf{x}_1,\mathbf{y}_1,\mathbf{o}_1,\ldots,\mathbf{a}_n,\mathbf{x}_n)^{-1})\\
&\qquad\qquad (\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n,x_n)\mid x_0=x_1=\cdots=x_n\})=1\\
&(\mathsf{P}\circ(\mathbf{x}_0,\mathbf{y}_0,\mathbf{a}_1,\mathbf{x}_1,\mathbf{y}_1,\mathbf{o}_1,\ldots,\mathbf{a}_n,\mathbf{x}_n,\mathbf{y}_n)^{-1})\\
&\qquad\qquad (\{(x_0,y_0,a_1,x_1,y_1,o_1,\ldots,a_n,x_n,y_n)\mid x_0=x_1=\cdots=x_n\})=1
\end{aligned}
$$

$$(\mathsf{P} \circ (\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_1, \mathbf{x}_1, \mathbf{y}_1, \mathbf{o}_1, \ldots, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n, \mathbf{o}_n)^{-1})$$
$$(\{(x_0, y_0, a_1, x_1, y_1, o_1, \ldots, a_n, x_n, y_n, o_n) \mid x_0 = x_1 = \cdots = x_n\}) = 1.$$

It follows that, for all $n \in \mathbb{N}_0$,

$$\mathsf{P}(\{\omega \mid \mathbf{x}(\omega)(j) = \mathbf{x}(\omega)(j+1), \text{ for all } j = 0, \ldots, n\}) = 1.$$

Then, since P is countably additive,

$$\mathsf{P}(\{\omega \mid \mathbf{x}(\omega)(n) = \mathbf{x}(\omega)(n+1), \text{ for all } n \in \mathbb{N}_0\}) = 1.$$

That is, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. □

Now the discussion turns to Bayesian inference in the conditional case. The next result shows that Bayesian inference in the conditional case is a special case of stochastic filtering. This result is important because it shows that all the methods of Bayesian machine learning can be regarded as special cases of stochastic filtering. The essential condition needed to reduce stochastic filtering to Bayesian inference is that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ be constant-valued a.s.

In this setting, there are no actions and there is no transition model. A history is just a finite sequence of observations and filtering consists of a sequence of observation updates only. Also, in this setting, the observation model (for $\mathbf{y}$ given $\mathbf{x}$) is a sequence $\xi \triangleq (\xi_n)_{n \in \mathbb{N}}$, where

$$\xi_n : H_{n-1} \times X \times Y \to \mathcal{P}(O)$$

is a regular conditional distribution of $\mathbf{o}_n$ given $(\mathbf{h}_{n-1}, \mathbf{x}_n, \mathbf{y}_n)$, for all $n \in \mathbb{N}$.

Proposition 4.2.9. *(Bayesian inference in the conditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{Y})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes,*

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$$

the schema for $\mathbf{y}$ given $\mathbf{x}$,

$$(\xi_n : H_{n-1} \times X \times Y \to \mathcal{P}(O))_{n \in \mathbb{N}}$$

the observation model for $\mathbf{y}$ given $\mathbf{x}$, υ_O a σ-finite measure on $\mathcal{O}$, υ_X a σ-finite measure on $\mathcal{X}$, and υ_Y a σ-finite measure on $\mathcal{Y}$. Suppose that, for all $n \in \mathbb{N}_0$, there is a conditional density $\breve{\mu}_n : H_n \times X \to \mathcal{D}(Y)$ such that $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$ and that, for all $n \in \mathbb{N}$, there is a conditional density $\breve{\xi}_n : H_{n-1} \times X \times Y \to \mathcal{D}(O)$ such that $\xi_n = \breve{\xi}_n \cdot \upsilon_O$. Suppose also that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ are constant-valued a.s. Then the following hold.

1. *For all $n \in \mathbb{N}_0$,*

$$\mu_{n+1} = \lambda(h, o, x).\lambda y.\breve{\xi}_{n+1}(h, x, y)(o) * \lambda(h, o, x).\mu_n(h, x) \;\; \mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $h_{n+1} \triangleq (h_n, o_{n+1}) \in H_{n+1}$,

$$\lambda x.\mu_{n+1}(h_{n+1}, x) = \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * \lambda x.\mu_n(h_n, x),$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, *where* $\mathcal{L}((h_{n+1}, x_{n+1}))(N_{n+1}) = 0$.

3. *For all* $n \in \mathbb{N}_0$ *and* $h_{n+1} \triangleq (h_n, o_{n+1}) \in H_{n+1}$,

$$\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) = \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * \lambda x.\breve{\mu}_n(h_n, x) \quad \upsilon_Y\text{-a.e.},$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, *where* $\mathcal{L}((h_{n+1}, x_{n+1}))(N_{n+1}) = 0$.

Proof. 1. For all $n \in \mathbb{N}_0$, $\mu_n : H_n \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{x}_n, \mathbf{h}_n)$ and so

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{x}_n)) = \lambda\omega.\mu_n((\mathbf{h}_n, \mathbf{x}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Since $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s., it follows that

$$\lambda\omega.\mu_n((\mathbf{h}_n, \mathbf{x}_n)(\omega))(C) = \lambda\omega.\mu_n((\mathbf{h}_n, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.}$$

Since $\mathbf{y} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s., it follows that $\mathbf{1}_{\mathbf{y}_n^{-1}(C)} = \mathbf{1}_{\mathbf{y}_{n+1}^{-1}(C)}$ a.s., for all $C \in \mathcal{Y}$. Thus, since also $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s., it follows that

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{x}_n)) = \mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{x}_{n+1})) \text{ a.s.}$$

Hence

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{x}_{n+1})) = \lambda\omega.\mu_n((\mathbf{h}_n, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. That is, for all $n \in \mathbb{N}_0$, $\mu_n : H_n \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{x}_{n+1})$.

Next, ξ_{n+1} is an observation model; hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})) = \lambda\omega.\xi_{n+1}((\mathbf{h}_n, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})(\omega))(B) \text{ a.s.}$$

Also $\breve{\xi}_{n+1}$ is a regular conditional density.

Now consider the probability kernel

$$\lambda(h, o, x).\lambda y.\breve{\xi}_{n+1}(h, x, y)(o) * \lambda(h, o, x).\mu_n(h, x) : H_n \times O \times X \to \mathcal{P}(Y).$$

By Proposition A.12.8, this probability kernel is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{o}_{n+1}, \mathbf{x}_{n+1})$. That is,

$$\lambda(h, o, x).\lambda y.\breve{\xi}_{n+1}(h, x, y)(o) * \lambda(h, o, x).\mu_n(h, x) : H_{n+1} \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_{n+1}, \mathbf{x}_{n+1})$. Since μ_{n+1} is by definition a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_{n+1}, \mathbf{x}_{n+1})$, it follows from the uniqueness part of Proposition A.5.16 that

$$\mu_{n+1} = \lambda(h, o, x).\lambda y.\breve{\xi}_{n+1}(h, x, y)(o) * \lambda(h, o, x).\mu_n(h, x) \;\; \mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))\text{-a.e.}$$

2. Hence, for all $n \in \mathbb{N}$ and $h_{n+1} \triangleq (h_n, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\lambda x.\mu_{n+1}(h_{n+1}, x) \\
&= \lambda x.(\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * \mu_n(h_n, x)) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * \lambda x.\mu_n(h_n, x),
\end{aligned}$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$.

3. For this part, recall Definition A.3.6. For all $n \in \mathbb{N}$ and $h_{n+1} \triangleq (h_n, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) \cdot \upsilon_Y \\
&= \lambda x.\mu_{n+1}(h_{n+1}, x) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * \lambda x.\mu_n(h_n, x) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * (\lambda x.\breve{\mu}_n(h_n, x) \cdot \upsilon_Y) \\
&= (\lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * \lambda x.\breve{\mu}_n(h_n, x)) \cdot \upsilon_Y. \qquad \text{[Proposition A.3.10]}
\end{aligned}$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$. The result now follows by Proposition A.2.11. □

$$\mu_{n+1} = \lambda(h, o, x).\lambda y.\breve{\xi}_{n+1}(h, x, y)(o) * \lambda(h, o, x).\mu_n(h, x)$$

$$\lambda x.\mu_{n+1}(h_{n+1}, x) = \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * \lambda x.\mu_n(h_n, x)$$

$$\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) = \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1}) * \lambda x.\breve{\mu}_n(h_n, x)$$

Figure 4.14: Bayesian inference in the conditional case

The recurrence equations for Bayesian inference in the conditional case are given in Figures 4.14 and 4.15. They show that, in the conditional case, stochastic filtering is a generalization of Bayesian inference.

$$\lambda x.\mu_{n+1}(h_{n+1}, x) = \lambda x.\lambda B.\frac{\int_Y \mathbf{1}_B\, \lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1})\, d\mu_n(h_n, x)}{\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1})\, d\mu_n(h_n, x)}$$

$$\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) = \lambda x.\frac{\lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1})\ \breve{\mu}_n(h_n, x)}{\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, x, y)(o_{n+1})\ \breve{\mu}_n(h_n, x)\, d\upsilon_Y}$$

Figure 4.15: Explicit form of Bayesian inference in the conditional case

4.2.2 Functional Case

In this subsection, the constant-valued a.s. condition is significantly weakened but at the expense of slightly more complicated filter recurrence equations. Here is the key definition.

Definition 4.2.3. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(X, \mathcal{X})$ a standard Borel space. A stochastic process $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is *functional almost surely* if there exists a sequence $(\alpha_n : X \to X)_{n \in \mathbb{N}_0}$ of isomorphisms such that

$$\mathsf{P}(\{\omega \in \Omega \mid \mathbf{x}(\omega)(n) = \alpha_n(\mathbf{x}(\omega)(n+1)), \text{ for all } n \in \mathbb{N}_0\}) = 1.$$

It is said that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is functional a.s. *via* $(\alpha_n : X \to X)_{n \in \mathbb{N}_0}$.

In other words, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is functional almost surely if

$$\mathsf{P}(\{\omega \in \Omega \mid \mathbf{x}_n(\omega) = \alpha_n(\mathbf{x}_{n+1}(\omega)), \text{ for all } n \in \mathbb{N}_0\}) = 1.$$

Note that $\{\omega \in \Omega \mid \mathbf{x}_n(\omega) = \alpha_n(\mathbf{x}_{n+1}(\omega)), \text{ for all } n \in \mathbb{N}_0\}$ is measurable, that is, an event. To see this, consider the set $C \triangleq \{f \in X^{\mathbb{N}_0} \mid f(n) = \alpha_n(f(n+1)), \text{ for all } n \in \mathbb{N}_0\}$. According to Proposition A.4.2, C is measurable. Since $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is measurable, $\mathbf{x}^{-1}(C)$ is a measurable subset of Ω. But $\{\omega \in \Omega \mid \mathbf{x}_n(\omega) = \alpha_n(\mathbf{x}_{n+1}(\omega)), \text{ for all } n \in \mathbb{N}_0\} = \mathbf{x}^{-1}(C)$.

'Functional almost surely' is usually abbreviated to 'functional a.s.'. If a stochastic process is constant-valued a.s., then it is functional a.s. with each $\alpha_n : X \to X$ the identity function.

Here is a consequence of the functional a.s. assumption.

Proposition 4.2.10. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes, $n \in \mathbb{N}_0$, and $\mu_n : H_n \times X \to \mathcal{P}(Y)$ a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{x}_n)$. Suppose that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

and also that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is functional a.s. via $(\alpha_n : X \to X)_{n \in \mathbb{N}_0}$. Then

$$\lambda(h, a, x).\mu_n(h, \alpha_n(x)) : H_n \times A \times X \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Proof. Note that $\lambda(h, a, x).\mu_n(h, \alpha_n(x))$ is a probability kernel. By Proposition 4.2.1,

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)) = \lambda\omega.\lambda(h, a, x).\mu_n(h, x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Since $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is functional a.s. via $(\alpha_n : X \to X)_{n \in \mathbb{N}_0}$, it follows that

$$\lambda\omega.\lambda(h, a, x).\mu_n(h, x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)(\omega))(C) = \lambda\omega.\lambda(h, a, x).\mu_n(h, x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \alpha_n \circ \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Next, let $D \in \mathcal{H}_n \otimes \mathcal{A} \otimes \mathcal{X}$. Let $B \triangleq (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})^{-1}(D)$, so that $B \in \sigma((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$. Now consider the isomorphism $\lambda(h,a,x).(h,a,\alpha_n(x)) : H_n \times A \times X \to H_n \times A \times X$. Then $D' \triangleq \lambda(h,a,x).(h,a,\alpha_n(x))(D) \in \mathcal{H}_n \otimes \mathcal{A} \otimes \mathcal{X}$. Let $B' \triangleq (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)^{-1}(D')$, so that $B' \in \sigma((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n))$. For all $\omega \in \Omega$ satisfying $\mathbf{x}_n(\omega) = \alpha_n(\mathbf{x}_{n+1}(\omega))$,

$$\begin{aligned}
&\omega \in B \\
&\text{iff } (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega) \in D \\
&\text{iff } (\mathbf{h}_n, \mathbf{a}_{n+1}, \alpha_n \circ \mathbf{x}_{n+1})(\omega) \in \lambda(h,a,x).(h,a,\alpha_n(x))(D) \\
&\text{iff } (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)(\omega) \in \lambda(h,a,x).(h,a,\alpha_n(x))(D) \\
&\text{iff } \omega \in B'.
\end{aligned}$$

Hence $\mathbf{1}_B = \mathbf{1}_{B'}$ a.s, because $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is functional a.s. Then, for all $C \in \mathcal{Y}$,

$$\begin{aligned}
&\int_\Omega \mathbf{1}_B \, \mathbf{1}_{\mathbf{y}_n^{-1}(C)} \, d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{B'} \, \mathbf{1}_{\mathbf{y}_n^{-1}(C)} \, d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{B'} \, \lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_n)(\omega))(C) \, d\mathsf{P} \\
&\qquad\qquad [\lambda(h,a,x).\mu_n(h,x) \text{ is a regular conditional distribution}] \\
&= \int_\Omega \mathbf{1}_B \, \lambda\omega.\lambda(h,a,x).\mu_n(h,x)((\mathbf{h}_n, \mathbf{a}_{n+1}, \alpha_n \circ \mathbf{x}_{n+1})(\omega))(C) \, d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_B \, \lambda\omega.\lambda(h,a,x).\mu_n(h,\alpha_n(x))((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \, d\mathsf{P}.
\end{aligned}$$

Thus

$$\mathsf{P}(\mathbf{y}_n^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})) = \lambda\omega.\lambda(h,a,x).\mu_n(h,\alpha_n(x))((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. That is, $\lambda(h,a,x).\mu_n(h,\alpha_n(x)) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$. □

Next is the result giving the filter recurrence equations for schemas and empirical beliefs in the conditional case, assuming that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is functional a.s.

Proposition 4.2.11. *(Filter recurrence equations for the functional, conditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes,*

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$$

the schema for $\mathbf{y}$ given $\mathbf{x}$,

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n \in \mathbb{N}}$$

the transition model for $\mathbf{y}$ *given* $\mathbf{x}$*,*

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}$$

the observation model for $\mathbf{y}$ *given* $\mathbf{x}$*,* υ_O *a* σ*-finite measure on* $\mathcal{O}$*,* υ_X *a* σ*-finite measure on* $\mathcal{X}$*, and* υ_Y *a* σ*-finite measure on* $\mathcal{Y}$*. Suppose that, for all* $n \in \mathbb{N}_0$*, there is a conditional density* $\breve{\mu}_n : H_n \times X \to \mathcal{D}(Y)$ *such that* $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$ *and that, for all* $n \in \mathbb{N}$*, there is a conditional density* $\breve{\tau}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(Y)$ *such that* $\tau_n = \breve{\tau}_n \cdot \upsilon_Y$ *and a conditional density* $\breve{\xi}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(O)$ *such that* $\xi_n = \breve{\xi}_n \cdot \upsilon_O$*. Suppose also that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n,\mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$*, and that* $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ *is functional a.s. via* $(\alpha_n : X \to X)_{n\in\mathbb{N}_0}$*. Then the following hold.*

1. *For all* $n \in \mathbb{N}_0$*,*

$$\mu_{n+1} = \lambda(h,a,o,x).\lambda y.\breve{\xi}_{n+1}(h,a,x,y)(o) * \lambda(h,a,o,x).(\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1})(h,a,x) \quad \mathcal{L}((\mathbf{h}_{n+1},\mathbf{x}_{n+1}))\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$*,*

$$\lambda x.\mu_{n+1}(h_{n+1},x) = \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * \lambda x.(\mu_n(h_n,\alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n,a_{n+1},x,y)),$$

   except on $\{x \in X \mid (h_{n+1},x) \in N_{n+1}\}$*, where* $\mathcal{L}((\mathbf{h}_{n+1},\mathbf{x}_{n+1}))(N_{n+1}) = 0$*.*

3. *For all* $n \in \mathbb{N}_0$ *and* $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$*,*

$$\lambda x.\breve{\mu}_{n+1}(h_{n+1},x) = \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n,a_{n+1},x,y)(o_{n+1}) * \lambda x.(\breve{\mu}_n(h_n,\alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n,a_{n+1},x,y)) \quad \upsilon_Y\text{-a.e.},$$

   except on $\{x \in X \mid (h_{n+1},x) \in N_{n+1}\}$*, where* $\mathcal{L}((\mathbf{h}_{n+1},\mathbf{x}_{n+1}))(N_{n+1}) = 0$*.*

Proof. 1. Clearly, $\lambda(h,a,x).\breve{\mu}_n(h,\alpha_n(x)) : H_n \times A \times X \to \mathcal{D}(Y)$ is a conditional density, for all $n \in \mathbb{N}_0$. Also, for all $n \in \mathbb{N}_0$,

$$\lambda(h,a,x).\mu_n(h,\alpha_n(x)) = \lambda(h,a,x).\breve{\mu}_n(h,\alpha_n(x)) \cdot \upsilon_Y.$$

To see this, since $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$, it follows that

$$\begin{aligned}
&(\lambda(h,a,x).\breve{\mu}_n(h,\alpha_n(x)) \cdot \upsilon_Y)(h,a,x)(C) \\
&= \int_Y \mathbf{1}_C \, \lambda(h,a,x).\breve{\mu}_n(h,\alpha_n(x))(h,a,x) \, d\upsilon_Y \\
&= \int_Y \mathbf{1}_C \, \breve{\mu}_n(h,\alpha_n(x)) \, d\upsilon_Y \\
&= (\breve{\mu}_n \cdot \upsilon_Y)(h,\alpha_n(x))(C) \\
&= \mu_n(h,\alpha_n(x))(C) \\
&= \lambda(h,a,x).\mu_n(h,\alpha_n(x))(h,a,x)(C),
\end{aligned}$$

for all $(h, a, x) \in H_n \times A \times X$ and $C \in \mathcal{Y}$. Hence

$$\lambda(h, a, x).\mu_n(h, \alpha_n(x)) = \lambda(h, a, x).\breve{\mu}_n(h, \alpha_n(x)) \cdot \upsilon_Y.$$

By Proposition 4.2.10, for all $n \in \mathbb{N}_0$, $\lambda(h, a, x).\mu_n(h, \alpha_n(x)) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Now, for all $n \in \mathbb{N}_0$, since τ_{n+1} is a regular conditional distribution,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)) = \lambda\omega.\tau_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Proposition A.7.15 now shows that, for all $C \in \mathcal{Y}$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})) = \lambda\omega.(\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \odot \tau_{n+1})((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.}$$

That is, for all $n \in \mathbb{N}_0$, $\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \odot \tau_{n+1} : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Furthermore,

$$\begin{aligned}
&\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \odot \tau_{n+1} \\
&= (\lambda(h, a, x).\breve{\mu}_n(h, \alpha_n(x)) \cdot \upsilon_Y) \odot (\breve{\tau}_{n+1} \cdot \upsilon_Y) \\
&= (\lambda(h, a, x).\breve{\mu}_n(h, \alpha_n(x)) \odot \breve{\tau}_{n+1}) \cdot \upsilon_Y. && \text{[Proposition A.3.8]}
\end{aligned}$$

Hence $\lambda(h, a, x).\breve{\mu}_n(h, \alpha_n(x)) \odot \breve{\tau}_{n+1}$ is a regular conditional density of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Next, ξ_{n+1} is an observation model; hence, for all $B \in \mathcal{O}$,

$$\mathsf{P}(\mathbf{o}_{n+1}^{-1}(B) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})) = \lambda\omega.\xi_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_{n+1})(\omega))(B).$$

Also $\breve{\xi}_{n+1}$ is a regular conditional density.

Now consider the probability kernel

$$\lambda(h, a, o, x).\lambda y.\breve{\xi}_{n+1}(h, a, x, y)(o) \ * \ \lambda(h, a, o, x).(\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \ \odot \ \tau_{n+1})(h, a, x) : H_n \times A \times O \times X \to \mathcal{P}(Y).$$

By Proposition A.12.8, this probability kernel is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{o}_{n+1}, \mathbf{x}_{n+1})$. Hence

$$\lambda(h, a, o, x).\lambda y.\breve{\xi}_{n+1}(h, a, x, y)(o) \ * \ \lambda(h, a, o, x).(\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \ \odot \ \tau_{n+1})(h, a, x) : H_{n+1} \to \mathcal{P}(Y)$$

is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_{n+1}, \mathbf{x}_{n+1})$. Since μ_{n+1} is by definition a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_{n+1}, \mathbf{x}_{n+1})$, it follows from the uniqueness part of Proposition A.5.16 that, for all $n \in \mathbb{N}_0$,

$$\mu_{n+1} = \lambda(h, a, o, x).\lambda y.\breve{\xi}_{n+1}(h, a, x, y)(o) * \lambda(h, a, o, x).(\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \odot \tau_{n+1})(h, a, x) \quad \mathcal{L}((\mathbf{h}_{n+1}, \mathbf{h}_{n+1}))\text{-a.e.}$$

2. Hence, for all $n \in \mathbb{N}$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\lambda x.\mu_{n+1}(h_{n+1}, x) \\
&= \lambda x.(\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) * (\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \odot \tau_{n+1})(h_n, a_{n+1}, x)) \\
&= \lambda x.(\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) * (\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) * \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)),
\end{aligned}$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$.

3. For this part, recall Definition A.3.6. For all $n \in \mathbb{N}$ and $h_{n+1} \triangleq (h_n, a_{n+1}, o_{n+1}) \in H_{n+1}$,

$$\begin{aligned}
&\lambda x.\breve{\mu}_{n+1}(h_{n+1}, x) \cdot \upsilon_Y \\
&= \lambda x.\mu_{n+1}(h_{n+1}, x) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) * \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) * \lambda x.((\breve{\mu}_n(h_n, \alpha_n(x)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y)) \\
&= \lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) * \lambda x.((\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y) \\
&\qquad\qquad \text{[Proposition A.3.8]} \\
&= (\lambda x.\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o_{n+1}) * \lambda x.(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \cdot \upsilon_Y, \\
&\qquad\qquad \text{[Proposition A.3.10]}
\end{aligned}$$

except on $\{x \in X \mid (h_{n+1}, x) \in N_{n+1}\}$, where $\mathcal{L}((\mathbf{h}_{n+1}, \mathbf{x}_{n+1}))(N_{n+1}) = 0$. The result now follows by Proposition A.2.11. □

Here is the observation model synthesis result for the functional case.

Proposition 4.2.12. *(Observation model synthesis for the functional, conditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(A, \mathcal{A})$ an action space, $(O, \mathcal{O})$ an observation space, $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ standard Borel spaces, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ an observation process, $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ and $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ stochastic processes,*

$$(\zeta_n : H_{n-1} \times A \times X \to \mathcal{P}(O))_{n \in \mathbb{N}}$$

the observation model for $\mathbf{x}$,

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$$

the schema for $\mathbf{y}$ given $\mathbf{x}$,

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n \in \mathbb{N}}$$

the transition model for $\mathbf{y}$ given $\mathbf{x}$,

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n \in \mathbb{N}}$$

the observation model for $\mathbf{y}$ given $\mathbf{x}$, υ_O a σ-finite measure on $\mathcal{O}$, υ_X a σ-finite measure on $\mathcal{X}$, and υ_Y a σ-finite measure on $\mathcal{Y}$. Suppose that, for all $n \in \mathbb{N}_0$, there is a conditional

density $\breve{\mu}_n : H_n \times X \to \mathcal{D}(Y)$ *such that* $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$. *Suppose that, for all* $n \in \mathbb{N}$, *there is a conditional density* $\breve{\zeta}_n : H_{n-1} \times A \times X \to \mathcal{D}(O)$ *such that* $\zeta_n = \breve{\zeta}_n \cdot \upsilon_O$, *a conditional density* $\breve{\tau}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(Y)$ *such that* $\tau_n = \breve{\tau}_n \cdot \upsilon_Y$, *and a conditional density* $\breve{\xi}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(O)$ *such that* $\xi_n = \breve{\xi}_n \cdot \upsilon_O$. *Suppose also that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$, *and that* $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ *is functional a.s. via* $(\alpha_n : X \to X)_{n \in \mathbb{N}_0}$. *Then the following hold.*

1. *For all* $n \in \mathbb{N}_0$,

$$\zeta_{n+1} = (\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \odot \tau_{n+1}) \odot \xi_{n+1} \quad \mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$*-almost all* $h_n \in H_n$, $a_{n+1} \in A$, *and* $x \in X$,

$$\zeta_{n+1}(h_n, a_{n+1}, x) = (\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y).$$

3. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$*-almost all* $h_n \in H_n$ $a_{n+1} \in A$, *and* $x \in X$,

$$\breve{\zeta}_{n+1}(h_n, a_{n+1}, x) =$$
$$(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \quad \upsilon_O\text{-a.e.}$$

4. *For all* $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$*-almost all* $h_n \in H_n$, $a_{n+1} \in A$, *and* $x \in X$, *and all* $C \in \mathcal{O}$,

$$\zeta_{n+1}(h_n, a_{n+1}, x)(C) =$$
$$\int_Y \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y)(C)\ d(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).$$

5. *For all* $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$*-almost all* $h_n \in H_n$, $a_{n+1} \in A$, *and* $x \in X$, *and* υ_O*-almost all* $o \in O$,

$$\breve{\zeta}_{n+1}(h_n, a_{n+1}, x)(o) =$$
$$\int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ d(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).$$

Proof. 1. By Proposition 4.2.10, for all $n \in \mathbb{N}_0$, $\lambda(h, a, x).\mu_n(h, \alpha_n(x)) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$. Since τ_{n+1} is a regular conditional distribution, for all $n \in \mathbb{N}_0$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)) = \lambda\omega.\tau_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Proposition A.7.15 now shows that, for all $C \in \mathcal{Y}$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})) =$$
$$\lambda\omega.(\lambda(h, a, x).\mu_n(h, \alpha_n(x)) \odot \tau_{n+1})((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.}$$

That is, for all $n \in \mathbb{N}_0$, $\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1} : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

Then, by Proposition A.7.15, $(\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1}) \odot \xi_{n+1} : H_n \times A \times X \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$, for all $n \in \mathbb{N}_0$. However, from Definition 4.1.2, $\zeta_{n+1} : H_n \times A \times X \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$, for all $n \in \mathbb{N}_0$. The result now follows from the uniqueness part of Proposition A.5.16.

2. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1)})$-almost all $h_n \in H_n$, $a_{n+1} \in A$, and $x \in X$,

$$\begin{aligned}
&\zeta_{n+1}(h_n, a_{n+1}, x) \\
&= ((\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1}) \odot \xi_{n+1})(h_n, a_{n+1}, x) \\
&= (\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1})(h_n, a_{n+1}, x) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= (\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y).
\end{aligned}$$

3. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$-almost all $h_n \in H_n$, $a_{n+1} \in A$, and $x \in X$,

$$\begin{aligned}
&\breve{\zeta}_{n+1}(h_n, a_{n+1}, x) \cdot \upsilon_O \\
&= \zeta_{n+1}(h_n, a_{n+1}, x) \\
&= (\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= ((\breve{\mu}_n(h_n, \alpha_n(x)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y)) \odot (\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&= ((\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&\qquad\qquad \text{[Proposition A.3.8]} \\
&= ((\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_O. \\
&\qquad\qquad \text{[Proposition A.3.8]}
\end{aligned}$$

The result now follows by Proposition A.2.11.

4. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$-almost all $h_n \in H_n$, $a_{n+1} \in A$, and $x \in X$, and all $C \in \mathcal{O}$,

$$\begin{aligned}
&\zeta_{n+1}(h_n, a_{n+1}, x)(C) \\
&= ((\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y))(C) \\
&= \int_Y \lambda y.\xi_{n+1}(h_n, a_{n+1}, x, y)(C)\ d(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).
\end{aligned}$$

5. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}))$-almost all $h_n \in H_n$, $a_{n+1} \in A$, and $x \in X$, and υ_O-almost all $o \in O$,

$$\begin{aligned}
&\breve{\zeta}_{n+1}(h_n, a_{n+1}, x)(o) \\
&= ((\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \odot \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y))(o) \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ (\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))\ d\upsilon_Y \\
&= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ \ d((\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y) \\
&\qquad\qquad \text{[Proposition A.3.3]}
\end{aligned}$$

$$= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ \ d((\breve{\mu}_n(h_n, \alpha_n(x)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y))$$

[Proposition A.3.8]

$$= \int_Y \lambda y.\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)\ d(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y)).$$

□

Next is the environment synthesis result for the functional case.

Proposition 4.2.13. *(Environment synthesis for the functional, conditional case) Let* $(\Omega, \mathfrak{S}, \mathsf{P})$ *be a probability space,* $(A, \mathcal{A})$ *an action space,* $(O, \mathcal{O})$ *an observation space,* $(X, \mathcal{X})$ *and* $(Y, \mathcal{Y})$ *standard Borel spaces,* $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ *an action process,* $\mathbf{o} : \Omega \to O^{\mathbb{N}}$ *an observation process,* $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ *and* $\mathbf{y} : \Omega \to Y^{\mathbb{N}_0}$ *stochastic processes,*

$$(\nu_n : H_n \to \mathcal{P}(X))_{n \in \mathbb{N}_0}$$

the schema for $\mathbf{x}$*,*

$$(\eta_n : H_{n-1} \times A \times X \to \mathcal{P}(X))_{n \in \mathbb{N}}$$

the transition model for $\mathbf{x}$*,*

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n \in \mathbb{N}_0}$$

the schema for $\mathbf{y}$ *given* $\mathbf{x}$*,*

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n \in \mathbb{N}}$$

the transition model for $\mathbf{y}$ *given* $\mathbf{x}$*,*

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n \in \mathbb{N}}$$

the observation model for $\mathbf{y}$ *given* $\mathbf{x}$*,* υ_O *a* σ*-finite measure on* $\mathcal{O}$*,* υ_X *a* σ*-finite measure on* $\mathcal{X}$*, and* υ_Y *a* σ*-finite measure on* $\mathcal{Y}$*. Let* $(\Xi_n : H_{n-1} \times A \to \mathcal{P}(O))_{n \in \mathbb{N}}$ *be the environment for* $\mathbf{a}$ *and* $\mathbf{o}$*. Suppose that, for all* $n \in \mathbb{N}_0$*, there is a conditional density* $\breve{\nu}_n : H_n \to \mathcal{D}(X)$ *such that* $\nu_n = \breve{\nu}_n \cdot \upsilon_X$*, and a conditional density* $\breve{\mu}_n : H_n \times X \to \mathcal{D}(Y)$ *such that* $\mu_n = \breve{\mu}_n \cdot \upsilon_Y$*. Suppose that, for all* $n \in \mathbb{N}$*, there is a conditional density* $\breve{\Xi}_n : H_{n-1} \times A \to \mathcal{D}(O)$ *such that* $\Xi_n = \breve{\Xi}_n \cdot \upsilon_O$*, a conditional density* $\breve{\eta}_n : H_{n-1} \times A \times X \to \mathcal{D}(X)$ *such that* $\eta_n = \breve{\eta}_n \cdot \upsilon_X$*, a conditional density* $\breve{\tau}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(Y)$ *such that* $\tau_n = \breve{\tau}_n \cdot \upsilon_Y$*, and a conditional density* $\breve{\xi}_n : H_{n-1} \times A \times X \times Y \to \mathcal{D}(O)$ *such that* $\xi_n = \breve{\xi}_n \cdot \upsilon_O$*. Suppose also that*

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{x}_n)$$

and

$$\sigma(\mathbf{a}_{n+1}) \underset{\sigma(\mathbf{h}_n, \mathbf{x}_n)}{\perp\!\!\!\perp} \sigma(\mathbf{y}_n),$$

for all $n \in \mathbb{N}_0$*, and that* $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ *is functional a.s. via* $(\alpha_n : X \to X)_{n \in \mathbb{N}_0}$ *Then the following hold.*

1. *For all* $n \in \mathbb{N}_0$,

$$\Xi_{n+1} = ((\lambda(h,a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1})) \odot \xi_{n+1} \quad \mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))\text{-a.e.}$$

2. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$,

$$\begin{aligned}&\Xi_{n+1}(h_n, a_{n+1}) = \\ &((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \odot \\ &\qquad\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y).\end{aligned}$$

3. *For all* $n \in \mathbb{N}_0$ *and* $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$,

$$\begin{aligned}&\breve{\Xi}_{n+1}(h_n, a_{n+1}) = \\ &((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \odot \\ &\qquad\qquad \lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \quad \upsilon_O\text{-a.e.}\end{aligned}$$

4. *For all* $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$, *and all* $C \in \mathcal{O}$,

$$\begin{aligned}&\Xi_{n+1}(h_n, a_{n+1})(C) = \\ &\int_{X \times Y} \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y)(C) \\ &d((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))).\end{aligned}$$

5. *For all* $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$*-almost all* $h_n \in H_n$ *and* $a_{n+1} \in A$, *and* υ_O*-almost all* $o \in O$,

$$\begin{aligned}&\breve{\Xi}_{n+1}(h_n, a_{n+1})(o) = \\ &\int_{X \times Y} \lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\ &d((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))).\end{aligned}$$

Proof. 1. By Proposition 4.1.1, $\lambda(h,a).\nu_n(h) : H_n \times A \to \mathcal{P}(X)$ is a regular conditional distribution of $\mathbf{x}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. Then, by Proposition A.7.15, for all $n \in \mathbb{N}_0$, $\lambda(h,a).\nu_n(h) \odot \eta_{n+1} : H_n \times A \to \mathcal{P}(X)$ is a regular conditional distribution of $\mathbf{x}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$.

By Proposition 4.2.10, for all $n \in \mathbb{N}_0$, $\lambda(h,a,x).\mu_n(h,\alpha_n(x)) : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_n$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$. Since τ_{n+1} is a regular conditional distribution, for all $n \in \mathbb{N}_0$,

$$\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)) = \lambda\omega.\tau_{n+1}((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1}, \mathbf{y}_n)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{Y}$. Proposition A.7.15 now shows that, for all $C \in \mathcal{Y}$,

$$\begin{aligned}&\mathsf{P}(\mathbf{y}_{n+1}^{-1}(C) \mid (\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})) = \\ &\qquad \lambda\omega.(\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1})((\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})(\omega))(C) \text{ a.s.}\end{aligned}$$

That is, for all $n \in \mathbb{N}_0$, $\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1} : H_n \times A \times X \to \mathcal{P}(Y)$ is a regular conditional distribution of $\mathbf{y}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1}, \mathbf{x}_{n+1})$.

By Proposition A.7.12, $(\lambda(h,a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1}) : H_n \times A \to \mathcal{P}(X \times Y)$ is a regular conditional distribution of $(\mathbf{x}_{n+1}, \mathbf{y}_{n+1})$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. Then, by Proposition A.7.15, $((\lambda(h,a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1})) \odot \xi_{n+1} : H_n \times A \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. However, from Definition 2.2.2, $\Xi_{n+1} : H_n \times A \to \mathcal{P}(O)$ is a regular conditional distribution of $\mathbf{o}_{n+1}$ given $(\mathbf{h}_n, \mathbf{a}_{n+1})$, for all $n \in \mathbb{N}_0$. The result now follows from the uniqueness part of Proposition A.5.16.

2. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$,

$$
\begin{aligned}
&\Xi_{n+1}(h_n, a_{n+1}) \\
&= (((\lambda(h,a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1})) \odot \xi_{n+1})(h_n, a_{n+1}) \\
&= ((\lambda(h,a).\nu_n(h) \odot \eta_{n+1}) \otimes (\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1}))(h_n, a_{n+1}) \odot \\
&= ((\lambda(h,a).\nu_n(h) \odot \eta_{n+1})(h_n, a_{n+1}) \otimes \lambda x.(\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1})(h_n, a_{n+1}, x)) \odot \\
&\qquad\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= ((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\lambda(h,a,x).\mu_n(h,\alpha_n(x)) \odot \tau_{n+1})(h_n, a_{n+1}, x)) \odot \\
&\qquad\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= ((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y).
\end{aligned}
$$

3. For all $n \in \mathbb{N}_0$ and $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$,

$$
\begin{aligned}
&\breve{\Xi}_{n+1}(h_n, a_{n+1}) \cdot \upsilon_O \\
&= \Xi_{n+1}(h_n, a_{n+1}) \\
&= ((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad\qquad \lambda(x,y).\xi_{n+1}(h_n, a_{n+1}, x, y) \\
&= (((\breve{\nu}_n(h_n) \cdot \upsilon_X) \odot (\lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x) \cdot \upsilon_X)) \otimes \\
&\quad \lambda x.((\breve{\mu}_n(h_n, \alpha_n(x)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y))) \odot \\
&\qquad\qquad (\lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \cdot \upsilon_X) \otimes \\
&\quad \lambda x.((\breve{\mu}_n(h_n, \alpha_n(x)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y))) \odot \\
&\qquad\qquad (\lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&\qquad\qquad \text{[Proposition A.3.8]} \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \cdot \upsilon_X) \otimes \\
&\quad \lambda x.((\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y)) \odot \\
&\qquad\qquad (\lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&\qquad\qquad \text{[Proposition A.3.8]} \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \cdot \\
&\qquad\qquad (\upsilon_X \otimes \upsilon_Y)) \odot (\lambda(x,y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_O) \\
&\qquad\qquad \text{[Proposition A.12.3]}
\end{aligned}
$$

$$= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \odot \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_O$$

[Proposition A.3.8]

The result now follows by Proposition A.2.11.

4. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$, and all $C \in \mathcal{O}$,

$$\begin{aligned}
&\Xi_{n+1}(h_n, a_{n+1})(C) \\
&= (((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad \lambda(x, y).\xi_{n+1}(h_n, a_{n+1}, x, y))(C) \\
&= \int_{X \times Y} \lambda(x, y).\xi_{n+1}(h_n, a_{n+1}, x, y)(C) \\
&\qquad d((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))).
\end{aligned}$$

5. For all $n \in \mathbb{N}_0$, $\mathcal{L}((\mathbf{h}_n, \mathbf{a}_{n+1}))$-almost all $h_n \in H_n$ and $a_{n+1} \in A$, and υ_O-almost all $o \in O$,

$$\begin{aligned}
&\breve{\Xi}_{n+1}(h_n, a_{n+1})(o) \\
&= (((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \odot \\
&\qquad \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y))(o) \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad ((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \\
&\qquad\qquad \lambda x.(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)))\, d(\upsilon_X \otimes \upsilon_Y) \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad d\,((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \otimes \\
&\qquad\qquad \lambda x.(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y))) \cdot (\upsilon_X \otimes \upsilon_Y) \\
&\qquad\qquad \text{[Proposition A.3.3]} \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad d\,(((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \cdot \upsilon_X) \otimes \\
&\qquad\qquad (\lambda x.(\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y)) \\
&\qquad\qquad \text{[Proposition A.12.3]} \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad d\,(((\breve{\nu}_n(h_n) \odot \lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x)) \cdot \upsilon_X) \otimes \\
&\qquad\qquad (\lambda x.((\breve{\mu}_n(h_n, \alpha_n(x)) \odot \lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y)) \cdot \upsilon_Y))) \\
&= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o) \\
&\qquad d\,(((\breve{\nu}_n(h_n) \cdot \upsilon_X) \odot (\lambda x.\breve{\eta}_{n+1}(h_n, a_{n+1}, x) \cdot \upsilon_X)) \otimes
\end{aligned}$$

$$\lambda x.((\breve{\mu}_n(h_n, \alpha_n(x)) \cdot \upsilon_Y) \odot (\lambda y.\breve{\tau}_{n+1}(h_n, a_{n+1}, x, y) \cdot \upsilon_Y)))$$

[Proposition A.3.8]

$$= \int_{X \times Y} \lambda(x, y).\breve{\xi}_{n+1}(h_n, a_{n+1}, x, y)(o)$$
$$d((\nu_n(h_n) \odot \lambda x.\eta_{n+1}(h_n, a_{n+1}, x)) \otimes \lambda x.(\mu_n(h_n, \alpha_n(x)) \odot \lambda y.\tau_{n+1}(h_n, a_{n+1}, x, y))).$$

□

4.3 Nonconditional Particle Filters

In many applications, it is necessary to approximate the filtering process by means of particle filters. The main idea of particle filters is to approximate distributions with mixtures of Dirac measures. Suppose that $\mu_n(h_n) : \mathcal{P}(X)$ is an empirical belief. Then a particle filter approximates $\mu_n(h_n)$ by a distribution of the form $\frac{1}{N}\sum_{i=1}^{N} \delta_{x^{(i)}}$, where $x^{(i)} \in X$, for $i = 1, \ldots, N$. Each $x^{(i)}$, for $i = 1, \ldots, N$ is called a particle. In addition, it is necessary to deal with the conditional case, where $\lambda x.\mu_n(h_n, x) : X \to \mathcal{P}(Y)$. Thus conditional particle filters are needed.

Particle filters enable efficient filtering even in cases where exact filtering does not lead to tractable mathematical expressions. They allow the accurate calculation of integrals with respect to the empirical belief. However, other kinds of reasoning may be impeded by the approximation by Dirac mixtures.

The algorithms for particle filtering are based directly on the theory for filtering (in particular, Propositions 4.1.2 and 4.2.3). There is one algorithm for the nonconditional case and one for the conditional case. Each algorithm is simple, essentially just a loop containing two sampling statements and a weight computation. The nonconditional algorithm is just the traditional bootstrap algorithm (with a subtle modification). It will be shown how it is possible to essentially derive the form of the algorithms from the theory. The Monte Carlo part of the algorithms depends fundamentally on the Strong Law of Large Numbers that is explained in Section A.5.

The discussion begins with the particle filter corresponding to Proposition 4.1.2. This is the case where the schema has the form

$$(\nu_n : H_n \to \mathcal{P}(X))_{n \in \mathbb{N}_0},$$

the transition model has the form

$$(\eta_n : H_{n-1} \times A \times X \to \mathcal{P}(X))_{n \in \mathbb{N}},$$

and the observation model has the form

$$(\zeta_n : H_{n-1} \times A \times X \to \mathcal{P}(O))_{n \in \mathbb{N}}.$$

(That is, it will be convenient to use here the notation introduced in Section 4.2 for the nonconditional case rather than that of Section 4.1.) The relevant recurrence equation from Proposition 4.1.2 is

$$\nu_n(h_n) = \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x)(o_n) * (\nu_{n-1}(h_{n-1}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x)).$$

Definition 4.3.1. A *particle* in X is just an element of X. A *particle family* is a finite indexed family of particles.

A particle family of N particles (that is, the index set is $\{1, \ldots, N\}$) at time n is an indexed family, usually denoted by $(x_n^{(i)})_{i=1}^N$. Particles in a particle family are typically not pairwise distinct.

A particle filter provides an approximation for $\nu_n(h_n) : \mathcal{P}(X)$. Suppose that the particle family at time n is $(x_n^{(i)})_{i=1}^N$. Then the Dirac mixture measure $\frac{1}{N}\sum_{i=1}^N \delta_{x_n^{(i)}}$ determined by $(x_n^{(i)})_{i=1}^N$ approximates $\nu_n(h_n)$. That is, for all $n \in \mathbb{N}$,

$$\nu_n(h_n) \approx \frac{1}{N}\sum_{i=1}^N \delta_{x_n^{(i)}}.$$

Thus, if $f : X \to \mathbb{R}$ is an integrable function, then

$$\int_X f \; d\nu_n(h_n) \approx \frac{1}{N}\sum_{i=1}^N f(x_n^{(i)}).$$

The following analysis provides motivation for the recursive step of the particle filter algorithm (for the nonconditional case). Suppose the current particle family is $(x_{n-1}^{(i)})_{i=1}^N$, so that $\nu_{n-1}(h_{n-1}) \approx \frac{1}{N}\sum_{i=1}^N \delta_{x_{n-1}^{(i)}}$. Then

$$\begin{aligned}
&\nu_n(h_n) \\
&= \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x)(o_n) \ast (\nu_{n-1}(h_{n-1}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x)) \\
&\approx \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x)(o_n) \ast ((\frac{1}{N}\sum_{i=1}^N \delta_{x_{n-1}^{(i)}}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x)) \\
&= \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x)(o_n) \ast \frac{1}{N}\sum_{i=1}^N \eta_n(h_{n-1}, a_n, x_{n-1}^{(i)}) \qquad \text{[Proposition A.3.5]} \\
&\approx \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x)(o_n) \ast \frac{1}{N}\sum_{i=1}^N \delta_{\overline{x}_n^{(i)}} \qquad \text{[Proposition A.5.4]} \\
&\qquad\qquad \text{[where } \overline{x}_n^{(i)} \sim \frac{1}{N}\sum_{i'=1}^N \eta_n(h_{n-1}, a_n, x_{n-1}^{(i')}) \text{, for } i = 1, \ldots, N] \\
&= \frac{\lambda B. \int_X \mathbf{1}_B \; \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x)(o_n) \; d(\frac{1}{N}\sum_{i=1}^N \delta_{\overline{x}_n^{(i)}})}{\int_X \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x)(o_n) \; d(\frac{1}{N}\sum_{i=1}^N \delta_{\overline{x}_n^{(i)}})} \\
&= \frac{\lambda B. \frac{1}{N}\sum_{i=1}^N \mathbf{1}_B(\overline{x}_n^{(i)}) \; \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)}{\frac{1}{N}\sum_{i'=1}^N \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i')})(o_n)} \\
&= \frac{\lambda B. \frac{1}{N}\sum_{i=1}^N \delta_{\overline{x}_n^{(i)}}(B) \; \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)}{\frac{1}{N}\sum_{i'=1}^N \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i')})(o_n)}
\end{aligned}$$

$$= \sum_{i=1}^{N} \frac{\breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)}{\sum_{i'=1}^{N} \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i')})(o_n)} \delta_{\overline{x}_n^{(i)}}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \delta_{x_n^{(i)}}, \qquad \text{[Proposition A.5.4]}$$

$$\text{[where } x_n^{(i)} \sim \sum_{i'=1}^{N} \frac{\breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i')})(o_n)}{\sum_{i''=1}^{N} \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i'')})(o_n)} \delta_{\overline{x}_n^{(i')}}, \text{ for } i = 1, \ldots, N\text{]}$$

as expected. It is apparent from this that the recursive step of the particle filter algorithm essentially consists of two sampling steps:

$$\overline{x}_n^{(i)} \sim \frac{1}{N} \sum_{i'=1}^{N} \eta_n(h_{n-1}, a_n, x_{n-1}^{(i')}), \text{ for } i = 1, \ldots, N$$

and

$$x_n^{(i)} \sim \sum_{i'=1}^{N} \frac{\breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i')})(o_n)}{\sum_{i''=1}^{N} \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i'')})(o_n)} \delta_{\overline{x}_n^{(i')}}, \text{ for } i = 1, \ldots, N.$$

Based on this analysis, the details of the algorithm are now presented. The initialization of the particle filter is given in Figure 4.16. This assumes there is given an initial empirical belief $\nu_0(())$. The recursive step is given in Figure 4.17. This assumes that there is given a transition model η and an observation model ζ. The function *ParticleFilter* in Figure 4.17 implements the filter recurrence equation in Proposition 4.1.2 applied to Dirac mixture measures. The algorithm in Figure 4.17 is the same as the algorithm given in [30] (apart from the more general transition and observation models used here). This algorithm is a slightly modified version of the traditional bootstrap filter [61] that is widely employed in the literature ([156, Table 4.3], for example). The traditional algorithm uses instead $\overline{x}_n^{(i)} \sim \eta_n(h_{n-1}, a_n, x_{n-1}^{(i)})$ in the first sampling step; the mixture distribution used here and in [30] is preferred because it fits exactly the above analysis.

function *InitializeParticleFilter* **returns** Initial particle family $(x_0^{(i)})_{i=1}^N$;

for $i := 1$ **to** N **do**
    sample $x_0^{(i)} \sim \nu_0(())$;

return $(x_0^{(i)})_{i=1}^N$;

Figure 4.16: Initialization of the particle filter for the nonconditional case

The occurrences of h_{n-1} in Figure 4.17 are to be interpreted in the following way: either suitable conditional independence assumptions apply and all three occurrences of this argument are missing (as in the case of state distributions, for example) or else the bounded summarization function $\beta_n : H_n \to B_n$ discussed in Section 4.2 is employed and

function *ParticleFilter*$((x_{n-1}^{(i)})_{i=1}^N, h_{n-1}, a_n, o_n)$
returns Particle family $(x_n^{(i)})_{i=1}^N$ at time n;
inputs: Particle family $(x_{n-1}^{(i)})_{i=1}^N$ at time $n-1$,
history h_{n-1} up to time $n-1$,
action a_n at time n,
observation o_n at time n;

for $i := 1$ **to** N **do**
sample $\overline{x}_n^{(i)} \sim \frac{1}{N}\sum_{i'=1}^N \eta_n(h_{n-1}, a_n, x_{n-1}^{(i')})$;
$\widetilde{w}_n^{(i)} := \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)$;

for $i := 1$ **to** N **do**
$w_n^{(i)} := \frac{\widetilde{w}_n^{(i)}}{\sum_{i'=1}^N \widetilde{w}_n^{(i')}}$;

for $i := 1$ **to** N **do**
sample $x_n^{(i)} \sim \sum_{i'=1}^N w_n^{(i')} \delta_{\overline{x}_n^{(i')}}$;

return $(x_n^{(i)})_{i=1}^N$;

Figure 4.17: Recursive step of the particle filter for the nonconditional case

$\beta_{n-1}(h_{n-1})$ is passed to the function *ParticleFilter* instead of h_{n-1}. Also, under a suitable conditional independence assumption, the argument a_n of $\breve{\zeta}_n$ may be missing.

As is apparent from Figure 4.17, maintaining a particle family is nothing more than sequential sampling from certain mixture models. Suppose that one wants to sample from a mixture measure $\sum_{i=1}^N c_i\mu_i : \mathcal{P}(X)$, where X is a measurable space, $\mu_i : \mathcal{P}(X)$ and $c_i \geq 0$, for $i = 1, \ldots, N$, and $\sum_{i=1}^N c_i = 1$. Consider the categorical distribution on $\{1, \ldots, N\}$, where i has probability mass c_i, for $i = 1, \ldots, N$. Then to sample from $\sum_{i=1}^N c_i\mu_i$, one first samples j from $\{1, \ldots, N\}$ using this categorical distribution and then samples from X using the distribution μ_j. Note that, if $\mu_i = \delta_{a_i}$, where $a_i \in X$, for $i = 1, \ldots, N$, then any sample will always be one of the a_i.

The next topic is concerned with the interaction of environment synthesis (Proposition 4.1.3) and the approximation of the empirical belief $\nu_n(h_n) : \mathcal{P}(X)$ by a particle family. In Part 2 of Proposition 4.1.3, the equation (using the notation of this section) giving the environment is

$$\Xi_n(h_{n-1}, a_n) = (\nu_{n-1}(h_{n-1}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x)) \odot \lambda x.\zeta_n(h_{n-1}, a_n, x).$$

Assume that $\nu_{n-1}(h_{n-1})$ is approximated using the Dirac mixture measure determined by a particle family, so that

$$\nu_{n-1}(h_{n-1}) \approx \frac{1}{N}\sum_{i=1}^N \delta_{x_{n-1}^{(i)}}.$$

Thus, by Proposition A.3.5,

$$\nu_{n-1}(h_{n-1}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x) \approx \frac{1}{N}\sum_{i=1}^{N} \eta_n(h_{n-1}, a_n, x_{n-1}^{(i)}).$$

It follows that

$$\Xi_n(h_{n-1}, a_n) \approx (\frac{1}{N}\sum_{i=1}^{N} \eta_n(h_{n-1}, a_n, x_{n-1}^{(i)})) \odot \lambda x.\zeta_n(h_{n-1}, a_n, x).$$

In terms of densities, this approximation has the form

$$\breve{\Xi}_n(h_{n-1}, a_n) \approx (\frac{1}{N}\sum_{i=1}^{N} \breve{\eta}_n(h_{n-1}, a_n, x_{n-1}^{(i)})) \odot \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x),$$

or, more explicitly,

$$\breve{\Xi}_n(h_{n-1}, a_n) \approx \lambda o. \int_X \lambda x.\breve{\zeta}_n(h_{n-1}, a_n, x)(o)\ \frac{1}{N}\sum_{i=1}^{N} \breve{\eta}_n(h_{n-1}, a_n, x_{n-1}^{(i)})\ d\upsilon_X.$$

Thus, given the approximation $\nu_{n-1}(h_{n-1}) \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{x_{n-1}^{(i)}}$, the transition model $\breve{\eta}_n$, and the observation model $\breve{\zeta}_n$, one can approximately compute the likelihood $\breve{\Xi}_n(h_{n-1}, a_n)(o_n)$ of the observation o_n at time n. The integral can itself be approximated using Monte Carlo integration by sampling from the density $\frac{1}{N}\sum_{i=1}^{N} \breve{\eta}_n(h_{n-1}, a_n, x_{n-1}^{(i)})$, by Proposition A.5.3. The approximation of the observation likelihood $\breve{\Xi}_n(h_{n-1}, a_n)(o_n)$ can be used, as discussed in Section 4.2, to adjust any parameters that need to be learned in the definitions of $\breve{\eta}_n$ and/or $\breve{\zeta}_n$.

Example 4.3.1. This example continues the discussion about function spaces in Example 3.4.5 by considering how a particle filter might be used to approximate an empirical belief about a function space. The setting is a schema having the form $(\mu_n : H_n \to \mathcal{P}(W^Z))_{n\in\mathbb{N}_0}$. Here, W is a finite set of classes for classification and Z an arbitrary set, so that W^Z is a function space. As explained in Example 3.4.5, instead of $(\mu_n)_{n\in\mathbb{N}_0}$, one works with a schema of the form $(\lambda h.(\nu_n(h) \circ p^{-1}) : H_n \to \mathcal{P}(W^Z))_{n\in\mathbb{N}_0}$, where $(\nu_n : H_n \to \mathcal{P}(W^{Z/\sim}))_{n\in\mathbb{N}_0}$ is a schema, $p \triangleq \lambda f.(f \circ \pi) : W^{Z/\sim} \to W^Z$, and $\pi : Z \to Z/\sim$ is the canonical surjection.

Let $u : W^Z \to \mathbb{R}$ be an integrable function that, for example, gives the utility for each $f \in W^Z$. Suppose that it is desired to compute the expected utility $\int_{W^Z} u\ d(\nu_n(h_n) \circ p^{-1})$. Let C be the set of functions in W^Z that are constant on each equivalence class in the partition of Z. Then $(\nu_n(h_n) \circ p^{-1})(C) = 1$. Thus the support of the probability measure $\nu_n(h_n) \circ p^{-1}$ is the set of piecewise-constant functions in W^Z that are based on the partition induced by $\sim$. Hence it is only the definition of u on C, rather than $W^Z \setminus C$, that is important.

Since

$$\int_{W^Z} u\ d(\nu_n(h_n) \circ p^{-1}) = \int_{W^{Z/\sim}} u \circ p\ d\nu_n(h_n),$$

the problem reduces to building a particle filter for $(\nu_n)_{n\in\mathbb{N}_0}$. In this circumstance of employing a particle filter, deconstructing $(\nu_n)_{n\in\mathbb{N}_0}$ is not usually necessary. Thus, let the transition model have the form

$$(\eta_n : H_{n-1} \times A \times W^{Z/\sim} \to \mathcal{P}(W^{Z/\sim}))_{n\in\mathbb{N}}$$

and the observation model have the form

$$(\zeta_n : H_{n-1} \times A \times W^{Z/\sim} \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

Then Figure 4.17 gives the particle filter algorithm for $(\nu_n)_{n\in\mathbb{N}_0}$. Suppose the particle family at time n is $(f_n^{(i)})_{i=1}^N$, where $f_n^{(i)} \in W^{Z/\sim}$, for $i = 1, \ldots, N$. The space $W^{Z/\sim}$ can be interpreted either as a finite product space or else as the set of all functions from the finite set of equivalence classes $Z/\sim$ to W. Then

$$\int_{W^Z} u \; d(\nu_n(h_n) \circ p^{-1}) = \int_{W^{Z/\sim}} u \circ p \; d\nu_n(h_n) \approx \frac{1}{N} \sum_{i=1}^{N} u(p(f_n^{(i)})).$$

Note that each $p(f_n^{(i)})$ can be interpreted as a function from Z to W that is constant on each equivalence class in $Z/\sim$.

As noted in Example 3.4.5, much depends on a good choice of the equivalence relation $\sim$. Towards this, suppose that there is a connection between $\sim$ and the transition and observation models, so that $\sim$ is essentially a 'parameter' of these models. Then the environment synthesis result (Proposition 4.1.3 can be used to find a good choice for $\sim$. As shown just above, according to Proposition 4.1.3, the likelihood $\breve{\Xi}_n(h_{n-1}, a_n)(o_n)$ of the observation o_n at time n can be approximately computed. This provides a basis for adjusting the equivalence relation: change the equivalence relation so that $\breve{\Xi}_n(h_{n-1}, a_n)(o_n)$ is increased. Ultimately, the only empirical information available to the agent comes from observations; hence observations should be used to improve the choice of the equivalence relation on Z.

Finally, in this section, the special case when $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. is considered. In this case, by Proposition 4.1.6, the transition model is such that every action is a no-op. The assumption that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. thus results in the algorithm given in Figure 4.18. Note carefully that, due to the particular form of the transition model η, for each $i \in \{1, \ldots, N\}$, there exists $i^* \in \{1, \ldots, N\}$ such that $x_n^{(i)} = x_{n-1}^{(i^*)}$.

In practice, this property of the transition model being a no-op can lead to degeneracy of the particle family. Thus a jittering transition model may be used instead [32]. Instead of keeping the parameter fixed, the transition model allows it to change slightly, thus avoiding the degeneracy. Commonly, the parameter space X is $\mathbb{R}^m$, for some $m \geq 1$. The transition model for the parameter space has the form $(\eta_n : X \to \mathcal{D}(X))_{n\in\mathbb{N}}$. In this model, η_n is defined, for all $n \in \mathbb{N}$, by

$$\eta_n = \lambda p.\mathcal{N}(p, \Sigma),$$

where Σ is a (diagonal) covariance matrix with suitably small diagonal arguments that may change over time according to some criterion if the jittering is adaptive. (Recall that

function *ParticleFilter*$((x_{n-1}^{(i)})_{i=1}^N, h_{n-1}, a_n, o_n)$
returns Particle family $(x_n^{(i)})_{i=1}^N$ at time n;
inputs: Particle family $(x_{n-1}^{(i)})_{i=1}^N$ at time $n-1$,
history h_{n-1} up to time $n-1$,
action a_n at time n,
observation o_n at time n;

for $i := 1$ **to** N **do**
sample $\overline{x}_n^{(i)} \sim \frac{1}{N}\sum_{i'=1}^N \delta_{x_{n-1}^{(i')}}$;
$\widetilde{w}_n^{(i)} := \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)$;

for $i := 1$ **to** N **do**
$w_n^{(i)} := \frac{\widetilde{w}_n^{(i)}}{\sum_{i'=1}^N \widetilde{w}_n^{(i')}}$;

for $i := 1$ **to** N **do**
sample $x_n^{(i)} \sim \sum_{i'=1}^N w_n^{(i')} \delta_{\overline{x}_n^{(i')}}$;

return $(x_n^{(i)})_{i=1}^N$;

Figure 4.18: Recursive step of the particle filter for the nonconditional case, where $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s.

$\mathcal{N}(\mu, \Sigma) : \mathcal{D}(\mathbb{R}^m)$ denotes the Gaussian density with mean μ and covariance matrix Σ.) More explicitly,

$$\eta_n = \lambda p.\lambda x.\frac{1}{(2\pi)^{m/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-p)^T\Sigma^{-1}(x-p)\right\}.$$

The corresponding change to the algorithm in Figure 4.18 is given in Figure 4.19 below.

4.4 Conditional Particle Filters

Now the discussion turns to the particle filter for conditional schemas. This is the case where the schema has the form

$$(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0},$$

the transition model has the form

$$(\tau_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(Y))_{n\in\mathbb{N}},$$

and the observation model has the form

$$(\xi_n : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

function $ParticleFilter((x_{n-1}^{(i)})_{i=1}^{N}, h_{n-1}, a_n, o_n)$
returns Particle family $(x_n^{(i)})_{i=1}^{N}$ at time n;
inputs: Particle family $(x_{n-1}^{(i)})_{i=1}^{N}$ at time $n-1$,
history h_{n-1} up to time $n-1$,
action a_n at time n,
observation o_n at time n;

for $i := 1$ **to** N **do**
sample $\overline{x}_n^{(i)} \sim \frac{1}{N}\sum_{i'=1}^{N} \mathcal{N}(x_{n-1}^{(i')}, \Sigma)$;
$\widetilde{w}_n^{(i)} := \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)$;

for $i := 1$ **to** N **do**
$w_n^{(i)} := \frac{\widetilde{w}_n^{(i)}}{\sum_{i'=1}^{N} \widetilde{w}_n^{(i')}}$;

for $i := 1$ **to** N **do**
sample $x_n^{(i)} \sim \sum_{i'=1}^{N} w_n^{(i')} \delta_{\overline{x}_n^{(i')}}$;

return $(x_n^{(i)})_{i=1}^{N}$;

Figure 4.19: Recursive step of particle filter for schemas having form $(\nu_n : H_n \to \mathcal{P}(X))_{n\in\mathbb{N}_0}$, where X is a space of parameters, there is a fixed but unknown parameter value, and the transition model for parameters is a jittering model

Also the assumption is made that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s.

In Figure 4.17, the particle family $(x_n^{(i)})_{i=1}^{N}$ approximates the probability measure which is the empirical belief $\nu_n(h_n) : \mathcal{P}(X)$. However, for a schema of the form $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$, the situation is more complicated since an associated empirical belief $\lambda x.\mu_n(h_n, x) : X \to \mathcal{P}(Y)$ is a probability kernel that is conditional on X. This means it is necessary to suitably generalize the concept of a particle family to a corresponding conditional concept.

Definition 4.4.1. A *conditional particle* (from X to Y) is a pair $(x, (y^{(j)})_{j=1}^{M})$, where x is a particle in X and $(y^{(j)})_{j=1}^{M}$ is a particle family in Y.

Definition 4.4.2. A *conditional particle family* (from X to Y) is an indexed family of the form $((x^{(i)}, (y^{(i,j)})_{j=1}^{M}))_{i=1}^{N}$, where each $(x^{(i)}, (y^{(i,j)})_{j=1}^{M})$ is a conditional particle.

So a conditional particle family is an indexed family of conditional particles. The empirical beliefs $\nu_n(h_n) : \mathcal{P}(X)$ and $\lambda x.\mu_n(h_n, x) : X \to \mathcal{P}(Y)$ together are approximated by a conditional particle family $((x_n^{(i)}, (y_n^{(i,j)})_{j=1}^{M}))_{i=1}^{N}$. More precisely,

$$\nu_n(h_n) \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{x_n^{(i)}},$$

and

$$\mu_n(h_n, x_n^{(i)}) \approx \frac{1}{M}\sum_{j=1}^{M} \delta_{y_n^{(i,j)}}, \text{ for } i = 1, \ldots, N.$$

Here is an analysis that motivates the conditional particle filter algorithm. Let the nonconditional particle family at time n be $(x_n^{(i)})_{i=1}^N$ and the conditional particle family at time $n-1$ be $((x_{n-1}^{(i)}, (y_{n-1}^{(i,j)})_{j=1}^M))_{i=1}^N$. Recall that, by Proposition 4.1.6, since $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s., for each $i \in \{1, \ldots, N\}$, there exists $i^* \in \{1, \ldots, N\}$ such that $x_n^{(i)} = x_{n-1}^{(i^*)}$. Thus, for $i = 1, \ldots, N$,

$$\begin{aligned}
&\mu_n(h_n, x_n^{(i)}) \\
&= \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, y)(o_n) * (\mu_{n-1}(h_{n-1}, x_n^{(i)}) \odot \lambda y.\tau_n(h_{n-1}, a_n, x_n^{(i)}, y)) \\
&= \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, y)(o_n) * (\mu_{n-1}(h_{n-1}, x_{n-1}^{(i^*)}) \odot \lambda y.\tau_n(h_{n-1}, a_n, x_n^{(i)}, y)) \\
&\approx \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, y)(o_n) * ((\frac{1}{M}\sum_{j=1}^{M} \delta_{y_{n-1}^{(i^*,j)}}) \odot \lambda y.\tau_n(h_{n-1}, a_n, x_n^{(i)}, y)) \\
&= \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, y)(o_n) \ * \ \frac{1}{M}\sum_{j=1}^{M} \tau_n(h_{n-1}, a_n, x_n^{(i)}, y_{n-1}^{(i^*,j)}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad [\text{Proposition A.3.5}] \\
&\approx \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, y)(o_n) \ * \ \frac{1}{M}\sum_{j=1}^{M} \delta_{\overline{y}_n^{(i,j)}} \qquad\qquad [\text{Proposition A.5.4}] \\
&\qquad\qquad [\text{where } \overline{y}_n^{(i,j)} \sim \frac{1}{M}\sum_{j'=1}^{M} \tau_n(h_{n-1}, a_n, x_n^{(i)}, y_{n-1}^{(i^*,j')}), \text{ for } j = 1, \ldots, M] \\
&= \frac{\lambda C. \int_Y \mathbf{1}_C \ \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, y)(o_n) \ d(\frac{1}{M}\sum_{j=1}^{M} \delta_{\overline{y}_n^{(i,j)}})}{\int_Y \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, y)(o_n) \ d(\frac{1}{M}\sum_{j=1}^{M} \delta_{\overline{y}_n^{(i,j)}})} \\
&= \frac{\lambda C. \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}_C(\overline{y}_n^{(i,j)}) \ \breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j)})(o_n)}{\frac{1}{M}\sum_{j'=1}^{M} \breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j')})(o_n)} \\
&= \frac{\lambda C. \frac{1}{M}\sum_{j=1}^{M} \delta_{\overline{y}_n^{(i,j)}}(C) \ \breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j)})(o_n)}{\frac{1}{M}\sum_{j'=1}^{M} \breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j')})(o_n)} \\
&= \sum_{j=1}^{M} \frac{\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j)})(o_n)}{\sum_{j'=1}^{M} \breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j')})(o_n)} \delta_{\overline{y}_n^{(i,j)}} \\
&\approx \frac{1}{M}\sum_{j=1}^{M} \delta_{y_n^{(i,j)}}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad [\text{Proposition A.5.4}] \\
&\qquad [\text{where } y_n^{(i,j)} \sim \sum_{j'=1}^{M} \frac{\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j')})(o_n)}{\sum_{j''=1}^{M} \breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j'')})(o_n)} \delta_{\overline{y}_n^{(i,j')}}, \text{ for } j = 1, \ldots, M]
\end{aligned}$$

It now follows from the above that

$$\begin{aligned}
&(\nu_n \otimes \mu_n)(h_n) \\
&= \nu_n(h_n) \otimes \lambda x.\mu_n(h_n, x) \\
&\approx (\frac{1}{N}\sum_{i=1}^{N} \delta_{x_n^{(i)}}) \otimes \lambda x.\mu_n(h_n, x) \\
&= \lambda A.\int_X (\lambda x.\int_Y \lambda y.\mathbf{1}_A(x,y)\ d\mu_n(h_n,x))\ d(\frac{1}{N}\sum_{i=1}^{N} \delta_{x_n^{(i)}}) \\
&= \lambda A.\frac{1}{N}\sum_{i=1}^{N}\int_Y \lambda y.\mathbf{1}_A(x_n^{(i)},y)\ d\mu_n(h_n,x_n^{(i)}) \\
&\approx \lambda A.\frac{1}{N}\sum_{i=1}^{N}\int_Y \lambda y.\mathbf{1}_A(x_n^{(i)},y)\ d(\frac{1}{M}\sum_{j=1}^{M} \delta_{y_n^{(i,j)}}) \\
&= \lambda A.\frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M} \mathbf{1}_A(x_n^{(i)},y_n^{(i,j)}) \\
&= \lambda A.\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \delta_{(x_n^{(i)},y_n^{(i,j)})}(A) \\
&= \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \delta_{(x_n^{(i)},y_n^{(i,j)})}
\end{aligned}$$

Thus

$$(\nu_n \otimes \mu_n)(h_n) \approx \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \delta_{(x_n^{(i)},y_n^{(i,j)})}.$$

It is also apparent from the analysis that, for $i = 1, \ldots, N$, the recursive step of the conditional particle filter algorithm essentially consists of two sampling steps:

$$\overline{y}_n^{(i,j)} \sim \frac{1}{M}\sum_{j'=1}^{M} \tau_n(h_{n-1}, a_n, x_n^{(i)}, y_{n-1}^{(i^*,j')}),\ \text{for } j = 1, \ldots, M$$

and

$$y_n^{(i,j)} \sim \sum_{j'=1}^{M} \frac{\breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j')})(o_n)}{\sum_{j''=1}^{M} \breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j'')})(o_n)} \delta_{\overline{y}_n^{(i,j')}},\ \text{for } j = 1, \ldots, M.$$

The initialization of a conditional particle filter for schemas having form $(\mu_n : H_n \times X \to \mathcal{P}(Y))_{n\in\mathbb{N}_0}$ is given in Figure 4.20. It is assumed that there is given an initial empirical belief $\lambda x.\mu_0(h_0, x) : X \to \mathcal{P}(Y)$. The input to the function *InitializeConditionalParticleFilter* is the initial particle family $(x_0^{(i)})_{i=1}^N$ that is returned by the function *InitializeParticleFilter*.

function *InitializeConditionalParticleFilter*$((x_0^{(i)})_{i=1}^N)$
returns Initial conditional particle family $((x_0^{(i)}, (y_0^{(i,j)})_{j=1}^M))_{i=1}^N$;
input: Initial particle family $(x_0^{(i)})_{i=1}^N$;

for $i := 1$ **to** N **do**
    for $j := 1$ **to** M **do**
        sample $y_0^{(i,j)} \sim \mu_0((), x_0^{(i)})$;

return $((x_0^{(i)}, (y_0^{(i,j)})_{j=1}^M))_{i=1}^N$;

Figure 4.20: Initialization of the particle filter for the conditional case

The recursive step of the particle filter for conditional schemas, where $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s., is given in Figure 4.21. This assumes that there is given a transition model τ and an observation model ξ. The function *ConditionalParticleFilter* in Figure 4.21 implements the filter recurrence equation in Proposition 4.2.3 for Dirac mixture measures on Y that are conditioned on particles in X. The output of *ConditionalParticleFilter* is a conditional particle family $((x_n^{(i)}, (y_n^{(i,j)})_{j=1}^M))_{i=1}^N$ such that $(x_n^{(i)})_{i=1}^N$ approximates the probability measure $\nu_n(h_n)$ and, for $i = 1, \ldots, N$, the particle family $(y_n^{(i,j)})_{j=1}^M$ that approximates the probability measure $\mu_n(h_n, x_n^{(i)})$.

If the algorithm in Figure 4.21 is used in conjunction with the nonconditional algorithm in Figure 4.18, then $x_n^{(i)} = x_{n-1}^{(i^*)}$, for some $i^* \in \{1, \ldots, N\}$. If it is used in conjunction with the nonconditional algorithm in Figure 4.19, then $x_n^{(i)} \sim \mathcal{N}(x_{n-1}^{(i^*)}, \Sigma)$.

Similar remarks to those made earlier about occurrences of h_{n-1} in Figure 4.17 also apply for Figure 4.21.

The particle filters in Figure 4.17 and Figure 4.21 together provide an approximation for $(\nu_n \otimes \mu_n)(h_n) : \mathcal{P}(X \times Y)$. Suppose that, from Figure 4.17, the particle family is $(x_n^{(i)})_{i=1}^N$ and, from Figure 4.21, the conditional particle family is $((x_n^{(i)}, (y_n^{(i,j)})_{j=1}^M))_{i=1}^N$. Then, for all $n \in \mathbb{N}$,

$$(\nu_n \otimes \mu_n)(h_n) \approx \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \delta_{(x_n^{(i)}, y_n^{(i,j)})}.$$

Thus, if $f : X \times Y \to \mathbb{R}$ is an integrable function, then

$$\int_{X \times Y} f \, d(\nu_n \otimes \mu_n)(h_n) \approx \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_n^{(i)}, y_n^{(i,j)}).$$

In addition, $\nu_n \odot \mu_n : H_n \to \mathcal{P}(Y)$ is the marginal probability kernel for $\nu_n \otimes \mu_n$ with

function *ConditionalParticleFilter*$((x_n^{(i)})_{i=1}^N, ((x_{n-1}^{(i)}, (y_{n-1}^{(i,j)})_{j=1}^M))_{i=1}^N, h_{n-1}, a_n, o_n)$
returns Conditional particle family $((x_n^{(i)}, (y_n^{(i,j)})_{j=1}^M))_{i=1}^N$ at time n;

inputs: Particle family $(x_n^{(i)})_{i=1}^N$ at time n,
conditional particle family $((x_{n-1}^{(i)}, (y_{n-1}^{(i,j)})_{j=1}^M))_{i=1}^N$ at time $n-1$,
history h_{n-1} up to time $n-1$,
action a_n at time n,
observation o_n at time n;

for $i := 1$ **to** N **do**
  for $j := 1$ **to** M **do**
    sample $\overline{y}_n^{(i,j)} \sim \frac{1}{M}\sum_{j'=1}^M \tau_n(h_{n-1}, a_n, x_n^{(i)}, y_{n-1}^{(i^*,j')})$;
    $\widetilde{w}_n^{(i,j)} := \breve{\xi}_n(h_{n-1}, a_n, x_n^{(i)}, \overline{y}_n^{(i,j)})(o_n)$;
  for $j := 1$ **to** M **do**
    $w_n^{(i,j)} := \frac{\widetilde{w}_n^{(i,j)}}{\sum_{j'=1}^M \widetilde{w}_n^{(i,j')}}$;
  for $j := 1$ **to** M **do**
    sample $y_n^{(i,j)} \sim \sum_{j'=1}^M w_n^{(i,j')} \delta_{\overline{y}_n^{(i,j')}}$;

return $((x_n^{(i)}, (y_n^{(i,j)})_{j=1}^M))_{i=1}^N$;

Figure 4.21: Recursive step of the particle filter for the conditional case, where $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s.

respect to Y. Thus, if $\pi_Y : X \times Y \to Y$ is the canonical projection, then

$$\begin{aligned}
&(\nu_n \odot \mu_n)(h_n) \\
&= (\nu_n \otimes \mu_n)(h_n) \circ \pi_Y^{-1} \\
&\approx (\frac{1}{NM}\sum_{i=1}^N \sum_{j=1}^M \delta_{(x_n^{(i)}, y_n^{(i,j)})}) \circ \pi_Y^{-1} \\
&= \frac{1}{NM}\sum_{i=1}^N \sum_{j=1}^M (\delta_{(x_n^{(i)}, y_n^{(i,j)})} \circ \pi_Y^{-1}) \\
&= \frac{1}{NM}\sum_{i=1}^N \sum_{j=1}^M \delta_{y_n^{(i,j)}}.
\end{aligned}$$

Thus, for all $n \in \mathbb{N}$,

$$(\nu_n \odot \mu_n)(h_n) \approx \frac{1}{NM}\sum_{i=1}^N \sum_{j=1}^M \delta_{y_n^{(i,j)}}.$$

Hence, if $f : Y \to \mathbb{R}$ is an integrable function, then

$$\int_Y f \; d(\nu_n \odot \mu_n)(h_n) \approx \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} f(y_n^{(i,j)}).$$

Note the asymmetry in the treatment of X and Y in the product space $X \times Y$, for which the order 'first X, then Y' is favoured. An alternative approach to that of this section is to not deconstruct the empirical belief on $X \times Y$ as has been done here. Thus $X \times Y$ is considered as a whole and empirical beliefs based on $X \times Y$ are filtered according to Figure 4.17 alone. The corresponding particle families have the form $((x_n^{(i)}, y_n^{(i)}))_{i=1}^N$ and so there are equal numbers of X and Y particles. In contrast, under the assumption that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s., the approach of this section produces, for each particle in X, M particles in Y. In some applications, this may be important for balancing the number of particles used to approximate the distribution on X and the various (conditional) distributions on Y.

The assignment statement

$$\widetilde{w}_n^{(i)} := \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)$$

in the algorithm in Figure 4.18 is now examined. The immediately preceding sampling statement in Figure 4.18 shows that $\overline{x}_n^{(i)} = x_{n-1}^{(i^*)}$, for some $i^* \in \{1, \ldots, N\}$. Hence it is required to evaluate $\breve{\zeta}_n(h_{n-1}, a_n, x_{n-1}^{(i^*)})(o_n)$, under the assumption that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. Towards this, according to Part 5 of Proposition 4.2.5,

$$\breve{\zeta}_n(h_{n-1}, a_n, x)(o) = \int_Y \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x, y)(o) \; d(\mu_{n-1}(h_{n-1}, x) \odot \lambda y.\tau_n(h_{n-1}, a_n, x, y)).$$

One can now use the approximation given by the conditional particle $(x_{n-1}^{(i^*)}, (y_{n-1}^{(i^*,j)})_{j=1}^M)$, so that

$$\mu_{n-1}(h_{n-1}, x_{n-1}^{(i^*)}) \approx \frac{1}{M} \sum_{j=1}^{M} \delta_{y_{n-1}^{(i^*,j)}}$$

and hence, by Proposition A.3.5,

$$\begin{aligned} &\mu_{n-1}(h_{n-1}, x_{n-1}^{(i^*)}) \odot \lambda y.\tau_n(h_{n-1}, a_n, x_{n-1}^{(i^*)}, y) \\ \approx &(\frac{1}{M} \sum_{j=1}^{M} \delta_{y_{n-1}^{(i^*,j)}}) \odot \lambda y.\tau_n(h_{n-1}, a_n, x_{n-1}^{(i^*)}, y) \\ = &\frac{1}{M} \sum_{j=1}^{M} \tau_n(h_{n-1}, a_n, x_{n-1}^{(i^*)}, y_{n-1}^{(i^*,j)}). \end{aligned}$$

Thus

$$\begin{aligned} &\breve{\zeta}_n(h_{n-1}, a_n, x_{n-1}^{(i^*)})(o_n) \approx \\ &\qquad \int_Y \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_{n-1}^{(i^*)}, y)(o_n) \; d\frac{1}{M} \sum_{j=1}^{M} \tau_n(h_{n-1}, a_n, x_{n-1}^{(i^*)}, y_{n-1}^{(i^*,j)}). \end{aligned}$$

It follows from this that the assignment

$$\widetilde{w}_n^{(i)} := \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)$$

in Figure 4.18 can be approximated by

$$\widetilde{w}_n^{(i)} := \int_Y \lambda y.\breve{\xi}_n(h_{n-1}, a_n, x_{n-1}^{(i^*)}, y)(o_n)\ d\frac{1}{M}\sum_{j=1}^{M} \tau_n(h_{n-1}, a_n, x_{n-1}^{(i^*)}, y_{n-1}^{(i^*,j)}).$$

The integral can itself be approximated using Monte Carlo integration by sampling from the probability measure $\frac{1}{M}\sum_{j=1}^{M} \tau_n(h_{n-1}, a_n, x_{n-1}^{(i^*)}, y_{n-1}^{(i^*,j)})$, by Proposition A.5.3.

Now the same assignment statement

$$\widetilde{w}_n^{(i)} := \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)$$

in Figure 4.19 is considered. The immediately preceding statement

$$\text{sample } \overline{x}_n^{(i)} \sim \tfrac{1}{N}\textstyle\sum_{i'=1}^{N} \mathcal{N}(x_{n-1}^{(i')}, \Sigma)$$

shows that $\overline{x}_n^{(i)} \sim \mathcal{N}(x_{n-1}^{(i^*)}, \Sigma)$, for some $i^* \in \{1, \ldots, N\}$. Suppose now that, whenever $\overline{x}_n^{(i)}$ closely approximates $x_{n-1}^{(i^*)}$, it is true that $\mu_{n-1}(h_{n-1}, \overline{x}_n^{(i)})$ closely approximates $\mu_{n-1}(h_{n-1}, x_{n-1}^{(i^*)})$. Then

$$\widetilde{w}_n^{(i)} := \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n)$$

can be approximated by

$$\widetilde{w}_n^{(i)} := \int_Y \lambda y.\breve{\xi}_n(h_{n-1}, a_n, \overline{x}_n^{(i)}, y)(o_n)\ d\frac{1}{M}\sum_{j=1}^{M} \tau_n(h_{n-1}, a_n, \overline{x}_n^{(i)}, y_{n-1}^{(i^*,j)}).$$

Thus the approximation of $\overline{x}_n^{(i)}$ by $x_{n-1}^{(i^*)}$ introduces an additional approximation in the computation of the weights that may or may not be negligible depending on the application.

The final topic of this section is concerned with the interaction of environment synthesis (Proposition 4.2.4) and approximations by particle families. To facilitate adaptive behaviour of an agent, it is useful to be able to evaluate expressions of the form $\breve{\Xi}_{n+1}(h_n, a_{n+1})(o_n)$, under the assumption that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. According to Part 5 of Proposition 4.2.4, the relevant equation is

$$\breve{\Xi}_n(h_{n-1}, a_n)(o) =$$
$$\int_{X\times Y} \lambda(x, y).\breve{\xi}_n(h_{n-1}, a_n, x, y)(o)$$
$$d((\nu_{n-1}(h_{n-1}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x)) \otimes \lambda x.(\mu_{n-1}(h_{n-1}, x) \odot \lambda y.\tau_n(h_{n-1}, a_n, x, y))).$$

To evaluate $\breve{\Xi}_n(h_{n-1}, a_n)(o)$ by Monte Carlo integration, it is necessary to sample from the probability measure $(\nu_{n-1}(h_{n-1}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x)) \otimes \lambda x.(\mu_{n-1}(h_{n-1}, x) \odot$

$\lambda y.\tau_n(h_{n-1}, a_n, x, y)$). Towards this, assume that $\nu_{n-1}(h_{n-1})$ is approximated using the Dirac mixture measure determined by a particle family, so that

$$\nu_{n-1}(h_{n-1}) \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{x_{n-1}^{(i)}}$$

and, similarly,

$$(\mu_{n-1})(h_{n-1}, x_{n-1}^{(i)}) \approx \frac{1}{M}\sum_{j=1}^{M} \delta_{y_{n-1}^{(i,j)}}.$$

Then

$$\begin{aligned}
&(\nu_{n-1}(h_{n-1}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x)) \otimes \lambda x.(\mu_{n-1}(h_{n-1}, x) \odot \lambda y.\tau_n(h_{n-1}, a_n, x, y)) \\
\approx\ &((\frac{1}{N}\sum_{i=1}^{N} \delta_{x_{n-1}^{(i)}}) \odot \lambda x.\delta_x) \otimes \lambda x.(\mu_{n-1}(h_{n-1}, x) \odot \lambda y.\tau_n(h_{n-1}, a_n, x, y)) \\
&\qquad\qquad \text{[Proposition 4.1.6]} \\
=\ &(\frac{1}{N}\sum_{i=1}^{N} \delta_{x_{n-1}^{(i)}}) \otimes \lambda x.(\mu_{n-1}(h_{n-1}, x) \odot \lambda y.\tau_n(h_{n-1}, a_n, x, y)) \\
&\qquad\qquad \text{[Proposition A.3.5]} \\
=\ &\lambda A.\int_X \left(\lambda x.\int_Y \lambda y.\mathbf{1}_A(x, y)\ d(\mu_{n-1}(h_{n-1}, x) \odot \lambda y.\tau_n(h_{n-1}, a_n, x, y))\right)\ d(\frac{1}{N}\sum_{i=1}^{N} \delta_{x_{n-1}^{(i)}}) \\
=\ &\lambda A.\frac{1}{N}\sum_{i=1}^{N}\int_Y \lambda y.\mathbf{1}_A(x_{n-1}^{(i)}, y)\ d(\mu_{n-1}(h_{n-1}, x_{n-1}^{(i)}) \odot \lambda y.\tau_n(h_{n-1}, a_n, x_{n-1}^{(i)}, y)) \\
\approx\ &\lambda A.\frac{1}{N}\sum_{i=1}^{N}\int_Y \lambda y.\mathbf{1}_A(x_{n-1}^{(i)}, y)\ d((\frac{1}{M}\sum_{j=1}^{M} \delta_{y_{n-1}^{(i,j)}}) \odot \lambda y.\tau_n(h_{n-1}, a_n, x_{n-1}^{(i)}, y)) \\
=\ &\lambda A.\frac{1}{N}\sum_{i=1}^{N}\int_Y \lambda y.\mathbf{1}_A(x_{n-1}^{(i)}, y)\ d(\frac{1}{M}\sum_{j=1}^{M} \tau_n(h_{n-1}, a_n, x_{n-1}^{(i)}, y_{n-1}^{(i,j)})) \\
&\qquad\qquad \text{[Proposition A.3.5]} \\
=\ &\lambda A.\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\int_Y \lambda y.\mathbf{1}_A(x_{n-1}^{(i)}, y)\ d\tau_n(h_{n-1}, a_n, x_{n-1}^{(i)}, y_{n-1}^{(i,j)}) \\
=\ &\lambda A.\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} \tau_n(h_{n-1}, a_n, x_{n-1}^{(i)}, y_{n-1}^{(i,j)})(\{y \mid (x_{n-1}^{(i)}, y) \in A\}).
\end{aligned}$$

Consequently, to sample from $(\nu_{n-1}(h_{n-1}) \odot \lambda x.\eta_n(h_{n-1}, a_n, x)) \otimes \lambda x.(\mu_{n-1}(h_{n-1}, x) \odot \lambda y.\tau_n(h_{n-1}, a_n, x, y))$, one should sample i from the uniform distribution on $\{1, \ldots, N\}$ and j from the uniform distribution on $\{1, \ldots, M\}$, then sample $\overline{y}_n^{(i,j)}$ from $\tau_n(h_{n-1}, a_n, x_{n-1}^{(i)}, y_{n-1}^{(i,j)})$, and return $(x_{n-1}^{(i)}, \overline{y}_n^{(i,j)})$. From that, Proposition A.5.3 can be used to approximately evaluate the observation likelihood $\breve{\Xi}_n(h_{n-1}, a_n)(o)$. This can then be used, as discussed in

Section 4.2, to adjust any parameters that need to be learned in the definitions of $\breve{\tau}_n$ and/or $\breve{\xi}_n$.

4.5 Factored Nonconditional Particle Filters

This section shows how so-called factored particle filters can be used to filter in high dimensions.

The well-known fundamental difficulty with particle filtering in high dimensions is that two distinct probability measures in high-dimensional spaces are nearly mutually singular. (For example, in high dimensions, Gaussian distributions with the identity matrix as covariance matrix have nearly all their probability mass in a thin annulus around a hypersphere with radius $\sqrt{d}$, where d is the dimension of the space. See the Gaussian Annulus Theorem in [15] or Section 3.3.3 in [160]. It follows that two distinct Gaussian distributions in a high-dimensional space are nearly mutually singular.) In particular, in high dimensions, the distribution obtained after a transition update and the distribution obtained after the subsequent observation update are nearly mutually singular. As a consequence, the particle family obtained by resampling gives a poor approximation of the distribution obtained from the observation update. Typically, what happens is that one particle from the transition update has (normalized) weight very nearly equal to one and so the resampled particle family degenerates to a single particle. See the discussion about this in [148], for example. Intuitively, the problem could be fixed with a large enough particle family. Unfortunately, it has been shown that to avoid degeneracy the size of the particle family has to be at least exponential in the problem size. More precisely, the size of the particle family must be exponential in the variance of the observation log likelihood, which depends not only on the state dimension but also on the distribution after the state transition and the number and character of observations. Simulations confirm this result. For the details, see [10], [12], and [148].

In spite of this difficulty, it is often possible to exploit the structure in the form of spatial locality of a particular high-dimensional problem to filter effectively. Consider the case of filtering epidemics. Let the state space for an epidemic be $\prod_{i=1}^{m} Y_i$ and $\{C_1, \ldots, C_p\}$ a partition of the index set $\{1, \ldots, m\}$. Suppose, for $l = 1, \ldots, p$, the size of C_l is small and that observations are available for subspaces of the form $\prod_{i \in C_l} Y_i$. It may even be the case that each C_l is a singleton. Then, since each C_l is small, the degeneracy difficulties mentioned above do not occur for the observation update for each $\prod_{i \in C_l} Y_i$. In addition, an assumption needs to be made about the transition model. It is too much to expect there to be local transition models completely confined to each $\prod_{i \in C_l} Y_i$. If this were true, it would be possible to filter the entire space by independently filtering on each of the subspaces. But what is often true is that the domain of transition model for $\prod_{i \in C_l} Y_i$ depends only on a subset of the Y_j, where index j is a neighbour of, or at least close by, an index in C_j. The use of such local transition models introduces an approximation of the state distribution but, as the experiment results here indicate, the error can be surprisingly small even for large graphs. The resulting particle filter algorithm is called the factored particle filter and is similar to the local particle filter of [133, 134].

Putting the conditional particle filter and the factored particle filter together, the factored conditional particle filter is obtained. For the purposes of tracking states and

estimating parameters, the parameter part of this algorithm is the same as for the conditional particle filter. However, the state part of this algorithm is different to the conditional particle filter because it is now factored.

In summary, the starting point is the standard particle filter of Figure 4.17. With such a filter, it is possible to acquire empirical beliefs in low-dimensional spaces. One extension of the standard particle filter is the conditional particle filter of Figure 4.21; with such a filter, it is possible to acquire empirical beliefs and estimate parameters in low-dimensional spaces. A different extension of the standard filter is the factored particle filter of Figure 4.27 below; with such a filter, it is possible to acquire empirical beliefs in high-dimensional spaces. The factored conditional particle filter of Figure 4.32 below combines the notions of a conditional particle filter and a factored particle filter, and makes it possible to acquire empirical beliefs and estimate parameters in high-dimensional spaces.

The presentation of factored particle filtering starts as usual with the simpler nonconditional case and then, in the next section, presents the conditional case, including parameter estimation.

The setting for this section is where the schema has the form

$$(\nu_n : H_n \to \mathcal{P}(\prod_{i=1}^{m} X_i))_{n\in\mathbb{N}_0},$$

the transition model has the form

$$(\eta_n : H_{n-1} \times A \times \prod_{i=1}^{m} X_i \to \mathcal{P}(\prod_{i=1}^{m} X_i))_{n\in\mathbb{N}},$$

and the observation model has the form

$$(\zeta_n : H_{n-1} \times A \times \prod_{i=1}^{m} X_i \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

Thus, in this section, attention is restricted to filtering in product spaces; indeed, the primary interest is in high-dimensional product spaces.

Consider the lattice of all partitions of the index set $\{1,\ldots,m\}$ for the product space $\prod_{i=1}^{m} X_i$. Each element of a partition is called a cluster. The partial order $\leq$ on this lattice is defined by $\mathfrak{A} \leq \mathfrak{B}$ if, for each cluster $C \in \mathfrak{A}$, there exists a cluster $D \in \mathfrak{B}$ such that $C \subseteq D$. It is said that $\mathfrak{A}$ is finer than $\mathfrak{B}$ or, equivalently, $\mathfrak{B}$ is coarser than $\mathfrak{A}$. The coarsest of all partitions is $\{\{1,\ldots,m\}\}$; the finest of all partitions is $\{\{1\},\ldots,\{m\}\}$.

The basic idea of factored particle filters is choose a suitable partition of the index set $\{1,\ldots,m\}$ and create particle filters on the subspaces of $\prod_{i=1}^{m} X_i$ that are obtained by forming the product subspace associated with each of the clusters in the partition. Thus let $\{C_1,\ldots,C_p\}$ be a partition of the index set $\{1,\ldots,m\}$. By permuting indices if necessary, it can be assumed that $\prod_{l=1}^{p}\prod_{i\in C_l} X_i$ can be identified with $\prod_{i=1}^{m} X_i$. For $l = 1,\ldots,p$, x_{C_l} denotes a typical element of $\prod_{i\in C_l} X_i$. Similarly, $(x_{C_1},\ldots,x_{C_p})$ denotes a typical element of $\prod_{i=1}^{m} X_i$.

The observation space corresponding to the lth cluster is O_l, for $l = 1,\ldots,p$. Thus $O = \prod_{l=1}^{p} O_l$. An observation for the lth cluster is denoted by $o^{(l)} \in O_l$. A history for all

the clusters together at time n is

$$h_n \triangleq (a_1, (o_1^{(1)}, \ldots, o_1^{(p)}), a_2, (o_2^{(1)}, \ldots, o_2^{(p)}), \ldots, a_n, (o_n^{(1)}, \ldots, o_n^{(p)})),$$

while a history for the lth cluster at time n is

$$h_n^{(l)} \triangleq (a_1, o_1^{(l)}, a_2, o_2^{(l)}, \ldots, a_n, o_n^{(l)}).$$

Of course, actions are shared amongst the clusters. If H_n is the set of all histories at time n for all the clusters together, then $H_n^{(l)}$ denotes the set of all histories at time n for the lth cluster.

To get started, the schema, the transition model, and the observation model are assumed to have particular factorizable forms depending on the choice of partition $\{C_1, \ldots, C_p\}$. (The assumption for the transition model will be weakened somewhat below.) For $l = 1, \ldots, p$, let

$$(\nu_n^{(l)} : H_n^{(l)} \to \mathcal{P}(\prod_{i \in C_l} X_i))_{n \in \mathbb{N}_0},$$

be a schema for which

$$\nu_n = \lambda h. \bigotimes_{l=1}^{p} \nu_n^{(l)}(h^{(l)}),$$

for all $n \in \mathbb{N}_0$. Also, for $l = 1, \ldots, p$, suppose that

$$(\eta_n^{(l)} : H_{n-1}^{(l)} \times A \times \prod_{i \in C_l} X_i \to \mathcal{P}(\prod_{i \in C_l} X_i))_{n \in \mathbb{N}},$$

is a transition model for which

$$\eta_n = \lambda(h, a, (x_{C_1}, \ldots, x_{C_p})). \bigotimes_{l=1}^{p} \eta_n^{(l)}(h^{(l)}, a, x_{C_l}),$$

for all $n \in \mathbb{N}$. Similarly, for $l = 1, \ldots, p$, suppose that

$$(\zeta_n^{(l)} : H_{n-1}^{(l)} \times A \times \prod_{i \in C_l} X_i \to \mathcal{P}(O_l))_{n \in \mathbb{N}},$$

is an observation model for which

$$\zeta_n = \lambda(h, a, (x_{C_1}, \ldots, x_{C_p})). \bigotimes_{l=1}^{p} \zeta_n^{(l)}(h^{(l)}, a, x_{C_l}),$$

for all $n \in \mathbb{N}$. Thus there is a schema $\nu^{(l)}$, transition model $\eta^{(l)}$, and an observation model $\zeta^{(l)}$ for each $\prod_{i \in C_l} X_i$ that is used to define the schema ν, transition model η, and observation model ζ, respectively, for the entire product space $\prod_{i=1}^{m} X_i$.

The above setting is actually possible, that is, under certain circumstances, there do exist schemas, transition models, and observation models that satisfy the above factorization conditions, as Proposition 4.5.1 below shows. In each case of schema, transition model, and observation model, there is a conditional independence assumption that is sufficient to ensure the desired factorization holds.

Proposition 4.5.1. *(Factorization for the nonconditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $\{C_1, \ldots, C_p\}$ a partition of $\{1, \ldots, m\}$, $(A, \mathcal{A})$ an action space, $(O_l, \mathcal{O}_l)$ an observation space, for $l = 1, \ldots, p$, $(X_i, \mathcal{A}_i)$ a measurable space, for $i = 1, \ldots, m$, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, and $\mathbf{o}^{(l)} : \Omega \to O_l^{\mathbb{N}}$ an observation process and $\mathbf{x}^{(C_l)} : \Omega \to (\prod_{i \in C_l} X_i)^{\mathbb{N}_0}$ a stochastic process, for $l = 1, \ldots, p$.*

1. Let $(\nu_n^{(l)} : H_n^{(l)} \to \mathcal{P}(\prod_{i \in C_l} X_i))_{n \in \mathbb{N}_0}$ be the schema for $\mathbf{x}^{(C_l)}$, for $l = 1, \ldots, p$, and

$$\nu_n \triangleq \lambda h. \bigotimes_{l=1}^{p} \nu_n^{(l)}(h^{(l)}) : H_n \to \mathcal{P}(\prod_{i=1}^{m} X_i),$$

for all $n \in \mathbb{N}_0$. Suppose that

$$\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{x}_n^{(C_l)}}^{-1}(A_l) \mid \mathbf{h}_n) = \prod_{l=1}^{p} \mathsf{P}({\mathbf{x}_n^{(C_l)}}^{-1}(A_l) \mid \mathbf{h}_n^{(l)}) \text{ a.s.},$$

for all $A_l \in \bigotimes_{i \in C_l} \mathcal{A}_i$ and for $l = 1, \ldots, p$. Then $(\nu_n : H_n \to \mathcal{P}(\prod_{i=1}^{m} X_i))_{n \in \mathbb{N}_0}$ is the schema for $(\mathbf{x}^{(C_1)}, \ldots, \mathbf{x}^{(C_p)})$.

2. Let $(\eta_n^{(l)} : H_{n-1}^{(l)} \times A \times \prod_{i \in C_l} X_i \to \mathcal{P}(\prod_{i \in C_l} X_i))_{n \in \mathbb{N}}$ be the transition model for $\mathbf{x}^{(C_l)}$, for $l = 1, \ldots, p$, and

$$\eta_n \triangleq \lambda(h, a, (x_{C_1}, \ldots, x_{C_p})). \bigotimes_{l=1}^{p} \eta_n^{(l)}(h^{(l)}, a, x_{C_l}) : H_{n-1} \times A \times \prod_{i=1}^{m} X_i \to \mathcal{P}(\prod_{i=1}^{m} X_i),$$

for all $n \in \mathbb{N}$. Suppose that

$$\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{x}_n^{(C_l)}}^{-1}(A_l) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, (\mathbf{x}_{n-1}^{(C_1)}, \ldots, \mathbf{x}_{n-1}^{(C_p)}))) = \prod_{l=1}^{p} \mathsf{P}({\mathbf{x}_n^{(C_l)}}^{-1}(A_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_{n-1}^{(C_l)})) \text{ a.s.},$$

for all $A_l \in \bigotimes_{i \in C_l} \mathcal{A}_i$ and for $l = 1, \ldots, p$. Then $(\eta_n : H_{n-1} \times A \times \prod_{i=1}^{m} X_i \to \mathcal{P}(\prod_{i=1}^{m} X_i))_{n \in \mathbb{N}}$ is the transition model for $(\mathbf{x}^{(C_1)}, \ldots, \mathbf{x}^{(C_p)})$.

3. Let $(\zeta_n^{(l)} : H_{n-1}^{(l)} \times A \times \prod_{i \in C_l} X_i \to \mathcal{P}(O_l))_{n \in \mathbb{N}}$ be the observation model for $\mathbf{x}^{(C_l)}$, for $l = 1, \ldots, p$, and

$$\zeta_n \triangleq \lambda(h, a, (x_{C_1}, \ldots, x_{C_p})). \bigotimes_{l=1}^{p} \zeta_n^{(l)}(h^{(l)}, a, x_{C_l}) : H_{n-1} \times A \times \prod_{i=1}^{m} X_i \to \mathcal{P}(\prod_{l=1}^{p} O_l),$$

for all $n \in \mathbb{N}$. Suppose that

$$\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{o}_n^{(l)}}^{-1}(D_l) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, (\mathbf{x}_n^{(C_1)}, \ldots, \mathbf{x}_n^{(C_p)}))) = \prod_{l=1}^{p} \mathsf{P}({\mathbf{o}_n^{(l)}}^{-1}(D_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n^{(C_l)})) \text{ a.s.},$$

for all $D_l \in \mathcal{O}_l$ and for $l = 1, \ldots, p$. Then $(\zeta_n : H_{n-1} \times A \times \prod_{i=1}^{m} X_i \to \mathcal{P}(\prod_{l=1}^{p} O_l))_{n \in \mathbb{N}}$ is the observation model for $(\mathbf{x}^{(C_1)}, \ldots, \mathbf{x}^{(C_p)})$.

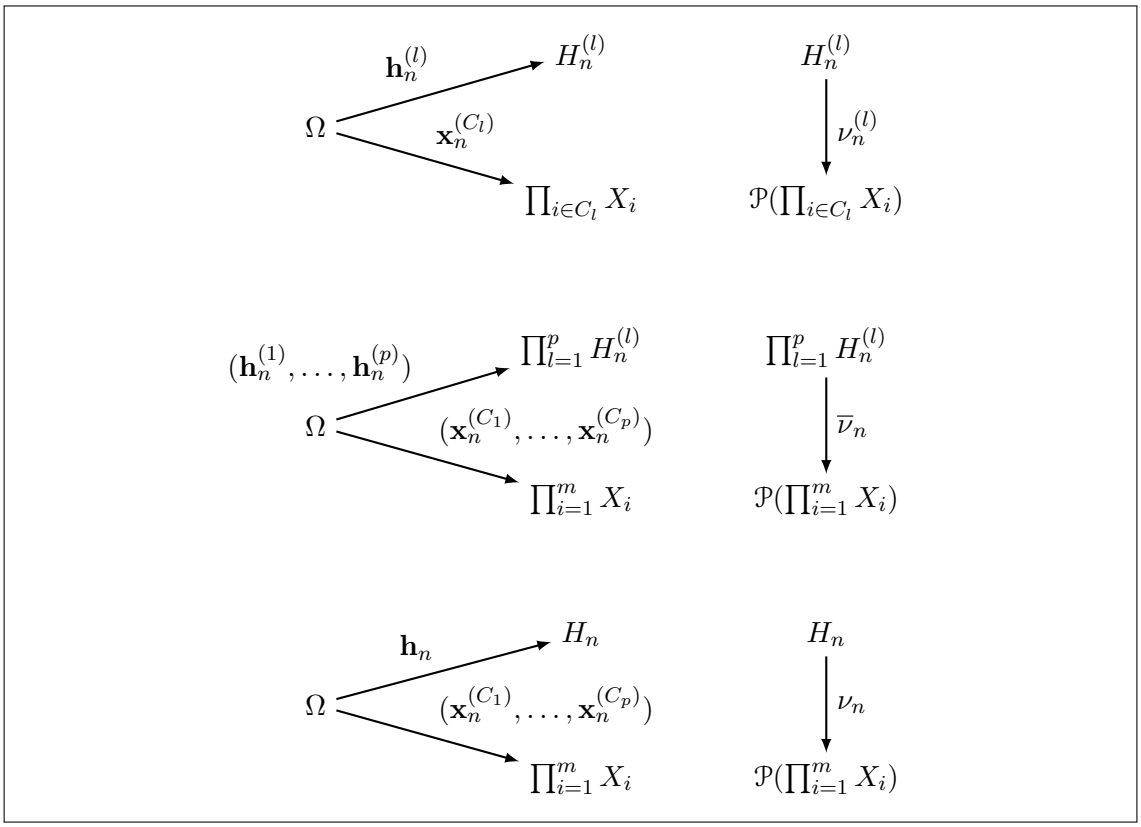

Figure 4.22: Setting for Part 1 of Proposition 4.5.1

Proof. 1. Since $(\nu_n^{(l)} : H_n^{(l)} \to \mathcal{P}(\prod_{i\in C_l} X_i))_{n\in\mathbb{N}_0}$ is the schema for $\mathbf{x}^{(C_l)}$, it follows that

$$\mathsf{P}({\mathbf{x}_n^{(C_l)}}^{-1}(A_l) \mid \mathbf{h}_n^{(l)}) = \lambda\omega.\nu_n^{(l)}(\mathbf{h}_n^{(l)}(\omega))(A_l) \text{ a.s.},$$

for all $A_l \in \bigotimes_{i\in C_l} \mathcal{A}_i$ and for $l = 1,\ldots,p$. Also, since $\sigma(\mathbf{h}_n) = \sigma((\mathbf{h}_n^{(1)},\ldots,\mathbf{h}_n^{(p)}))$, it follows that

$$\mathsf{P}(\bigcap_{l=1}^p {\mathbf{x}_n^{(C_l)}}^{-1}(A_l) \mid (\mathbf{h}_n^{(1)},\ldots,\mathbf{h}_n^{(p)})) = \prod_{l=1}^p \mathsf{P}({\mathbf{x}_n^{(C_l)}}^{-1}(A_l) \mid \mathbf{h}_n^{(l)}) \text{ a.s.},$$

for all $A_l \in \bigotimes_{i\in C_l} \mathcal{A}_i$ and for $l = 1,\ldots,p$.

Let

$$\overline{\nu}_n \triangleq \lambda(h_1,\ldots,h_p).\bigotimes_{l=1}^p \nu_n^{(l)}(h_l) : \prod_{l=1}^p H_n^{(l)} \to \mathcal{P}(\prod_{i=1}^m X_i).$$

Then, for all $A \in \bigotimes_{i=1}^m \mathcal{A}_i$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}((\mathbf{x}_n^{(C_1)},\ldots,\mathbf{x}_n^{(C_p)})^{-1}(A) \mid \mathbf{h}_n) \\
&= \mathsf{P}((\mathbf{x}_n^{(C_1)},\ldots,\mathbf{x}_n^{(C_p)})^{-1}(A) \mid (\mathbf{h}_n^{(1)},\ldots,\mathbf{h}_n^{(p)})) \qquad [\sigma(\mathbf{h}_n) = \sigma((\mathbf{h}_n^{(1)},\ldots,\mathbf{h}_n^{(p)}))]
\end{aligned}$$

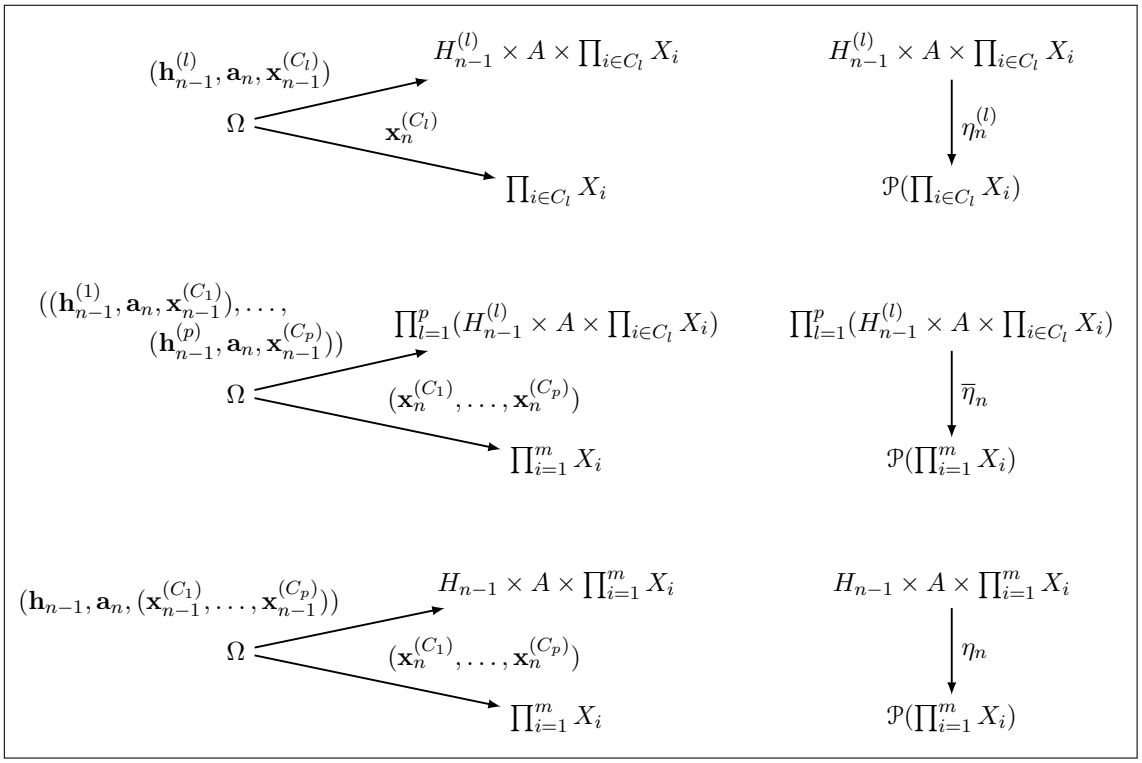

Figure 4.23: Setting for Part 2 of Proposition 4.5.1

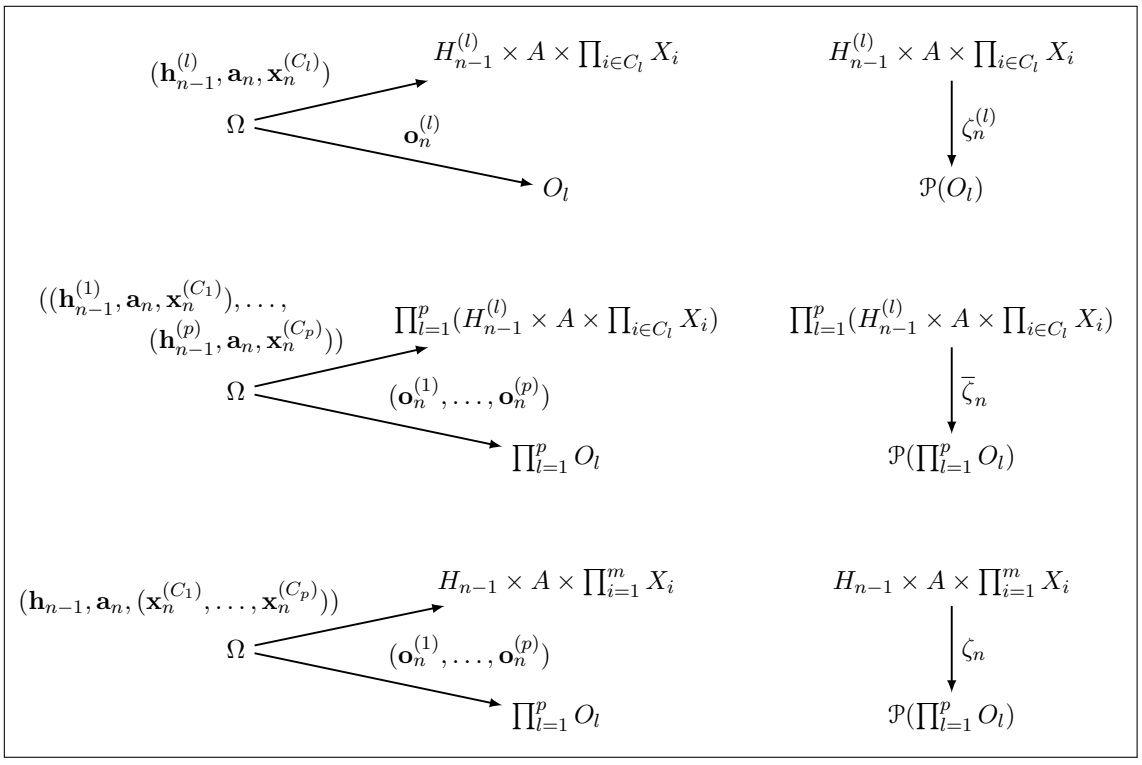

Figure 4.24: Setting for Part 3 of Proposition 4.5.1

$$
\begin{aligned}
&= \lambda\omega.\overline{\nu}_n((\mathbf{h}_n^{(1)},\ldots,\mathbf{h}_n^{(p)})(\omega))(A) && \text{[Proposition A.5.22]}\\
&= \lambda\omega.\bigotimes_{l=1}^{p}\nu_n^{(l)}(\mathbf{h}_n^{(l)}(\omega))(A)\\
&= \lambda\omega.\nu_n(\mathbf{h}_n(\omega))(A).
\end{aligned}
$$

That is, $(\nu_n : H_n \to \mathcal{P}(\prod_{i=1}^{m} X_i))_{n\in\mathbb{N}_0}$ is the schema for $(\mathbf{x}^{(C_1)},\ldots,\mathbf{x}^{(C_p)})$.

2. Since $(\eta_n^{(l)} : H_{n-1}^{(l)} \times A \times \prod_{i\in C_l} X_i \to \mathcal{P}(\prod_{i\in C_l} X_i))_{n\in\mathbb{N}}$ is the transition model for $\mathbf{x}^{(C_l)}$, it follows that

$$
\mathsf{P}(\mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid (\mathbf{h}_{n-1}^{(l)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_l)})) = \lambda\omega.\eta_n^{(l)}((\mathbf{h}_{n-1}^{(l)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_l)})(\omega))(A_l) \text{ a.s.},
$$

for all $A_l \in \bigotimes_{i\in C_l}\mathcal{A}_i$ and for $l = 1,\ldots,p$. Also, since

$$
\sigma((\mathbf{h}_{n-1},\mathbf{a}_n,(\mathbf{x}_{n-1}^{(C_1)},\ldots,\mathbf{x}_{n-1}^{(C_p)}))) = \sigma(((\mathbf{h}_{n-1}^{(1)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_1)}),\ldots,(\mathbf{h}_{n-1}^{(p)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_p)}))),
$$

it follows that

$$
\mathsf{P}(\bigcap_{l=1}^{p}\mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid ((\mathbf{h}_{n-1}^{(1)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_1)}),\ldots,(\mathbf{h}_{n-1}^{(p)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_p)}))) = \prod_{l=1}^{p}\mathsf{P}(\mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid (\mathbf{h}_{n-1}^{(l)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_l)})) \text{ a.s.},
$$

for all $A_l \in \bigotimes_{i\in C_l}\mathcal{A}_i$ and for $l = 1,\ldots,p$.

Let

$$
\overline{\eta}_n \triangleq \lambda((h_1,a_1,x_1),\ldots,(h_p,a_p,x_p)).\bigotimes_{l=1}^{p}\eta_n^{(l)}(h_l,a_l,x_l) : \prod_{l=1}^{p}(H_{n-1}^{(l)}\times A\times\prod_{i\in C_l}X_i) \to \mathcal{P}(\prod_{i=1}^{m}X_i).
$$

Then, for all $A \in \bigotimes_{i=1}^{m}\mathcal{A}_i$, P-almost surely,

$$
\begin{aligned}
&\mathsf{P}((\mathbf{x}_n^{(C_1)},\ldots,\mathbf{x}_n^{(C_p)})^{-1}(A) \mid (\mathbf{h}_{n-1},\mathbf{a}_n,(\mathbf{x}_{n-1}^{(C_1)},\ldots,\mathbf{x}_{n-1}^{(C_p)})))\\
&= \mathsf{P}((\mathbf{x}_n^{(C_1)},\ldots,\mathbf{x}_n^{(C_p)})^{-1}(A) \mid ((\mathbf{h}_{n-1}^{(1)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_1)}),\ldots,(\mathbf{h}_{n-1}^{(p)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_p)})))\\
&\qquad [\sigma((\mathbf{h}_{n-1},\mathbf{a}_n,(\mathbf{x}_{n-1}^{(C_1)},\ldots,\mathbf{x}_{n-1}^{(C_p)}))) = \sigma(((\mathbf{h}_{n-1}^{(1)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_1)}),\ldots,(\mathbf{h}_{n-1}^{(p)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_p)})))]\\
&= \lambda\omega.\overline{\eta}_n(((\mathbf{h}_{n-1}^{(1)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_1)}),\ldots,(\mathbf{h}_{n-1}^{(p)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_p)}))(\omega))(A) && \text{[Proposition A.5.22]}\\
&= \lambda\omega.\bigotimes_{l=1}^{p}\eta_n^{(l)}((\mathbf{h}_{n-1}^{(l)},\mathbf{a}_n,\mathbf{x}_{n-1}^{(C_l)})(\omega))(A)\\
&= \lambda\omega.\eta_n((\mathbf{h}_{n-1},\mathbf{a}_n,(\mathbf{x}_{n-1}^{(C_1)},\ldots,\mathbf{x}_{n-1}^{(C_p)}))(\omega))(A).
\end{aligned}
$$

That is, $(\eta_n : H_{n-1}\times A\times\prod_{i=1}^{m}X_i \to \mathcal{P}(\prod_{i=1}^{m}X_i))_{n\in\mathbb{N}}$ is the transition model for $(\mathbf{x}^{(C_1)},\ldots,\mathbf{x}^{(C_p)})$.

3. Since $(\zeta_n^{(l)} : H_{n-1}^{(l)} \times A \times \prod_{i\in C_l} X_i \to \mathcal{P}(O_l)_{n\in\mathbb{N}}$ is the observation model for $\mathbf{x}^{(C_l)}$, it follows that

$$\mathsf{P}({\mathbf{o}_n^{(l)}}^{-1}(D_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n^{(C_l)})) = \lambda\omega.\zeta_n^{(l)}((\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n^{(C_l)})(\omega))(D_l) \text{ a.s.},$$

for all $D_l \in \mathcal{O}_l$ and for $l = 1, \ldots, p$. Also, since

$$\sigma((\mathbf{h}_{n-1}, \mathbf{a}_n, (\mathbf{x}_n^{(C_1)}, \ldots, \mathbf{x}_n^{(C_p)}))) = \sigma(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n^{(C_p)}))),$$

it follows that

$$\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{o}_n^{(l)}}^{-1}(D_l) \mid ((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n^{(C_p)}))) = \prod_{l=1}^{p} \mathsf{P}({\mathbf{o}_n^{(l)}}^{-1}(D_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n^{(C_l)})) \text{ a.s.},$$

for all $D_l \in \mathcal{O}_l$ and for $l = 1, \ldots, p$.

Let

$$\overline{\zeta}_n \triangleq \lambda((h_1, a_1, x_1), \ldots, (h_p, a_p, x_p)).\bigotimes_{l=1}^{p} \zeta_n^{(l)}(h_l, a_l, x_l) : \prod_{l=1}^{p}(H_{n-1}^{(l)} \times A \times \prod_{i\in C_l} X_i) \to \mathcal{P}(\prod_{l=1}^{p} O_l).$$

Then, for all $D \in \bigotimes_{l=1}^{p} \mathcal{O}_l$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}((\mathbf{o}_n^{(1)}, \ldots, \mathbf{o}_n^{(p)})^{-1}(D) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, (\mathbf{x}_n^{(C_1)}, \ldots, \mathbf{x}_n^{(C_p)})))\\
&= \mathsf{P}((\mathbf{o}_n^{(1)}, \ldots, \mathbf{o}_n^{(p)})^{-1}(D) \mid ((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n^{(C_p)})))\\
&\qquad [\sigma((\mathbf{h}_{n-1}, \mathbf{a}_n, (\mathbf{x}_n^{(C_1)}, \ldots, \mathbf{x}_n^{(C_p)}))) = \sigma(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n^{(C_p)})))]\\
&= \lambda\omega.\overline{\zeta}_n(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n^{(C_p)}))(\omega))(D) \qquad \text{[Proposition A.5.22]}\\
&= \lambda\omega.\bigotimes_{l=1}^{p} \zeta_n^{(l)}((\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n^{(C_l)})(\omega))(D)\\
&= \lambda\omega.\zeta_n((\mathbf{h}_{n-1}, \mathbf{a}_n, (\mathbf{x}_n^{(C_1)}, \ldots, \mathbf{x}_n^{(C_p)}))(\omega))(D).
\end{aligned}$$

That is, $(\zeta_n : H_{n-1} \times A \times \prod_{i=1}^{m} X_i \to \mathcal{P}(\prod_{l=1}^{p} O_l))_{n\in\mathbb{N}}$ is the observation model for $(\mathbf{x}^{(C_1)}, \ldots, \mathbf{x}^{(C_p)})$. □

Next, particle families in each $\prod_{i\in C_l} X_i$ are related to particle families in $\prod_{i=1}^{m} X_i$. The notation $x^{(C_l,j)}$ denotes the jth particle of a particle family in $\prod_{i\in C_l} X_i$.

Definition 4.5.1. For $l = 1, \ldots, p$, let $(x^{(C_l,j_l)})_{j_l=1}^{N_j}$ be a particle family in $\prod_{i\in C_l} X_i$. Then the particle family in $\prod_{i=1}^{m} X_i$ *generated by* $(x^{(C_1,j_1)})_{j_1=1}^{N_1}, \ldots, (x^{(C_p,j_p)})_{j_p=1}^{N_p}$ is

$$((x^{(C_1,j_1)}, x^{(C_2,j_2)}, \ldots, x^{(C_p,j_p)}))_{j_1=1}^{N_1} \cdots_{j_p=1}^{N_p}.$$

$\prod_{i\in C_1} X_i:$	$x^{(C_1,1)}$	$x^{(C_1,2)}$	$x^{(C_1,3)}$	$\dots$	$\dots$	$x^{(C_1,N_1)}$	
$\prod_{i\in C_2} X_i:$	$x^{(C_2,1)}$	$x^{(C_2,2)}$	$x^{(C_2,3)}$	$\dots$	$x^{(C_2,N_2)}$		
$\prod_{i\in C_3} X_i:$	$x^{(C_3,1)}$	$x^{(C_3,2)}$	$x^{(C_3,3)}$	$\dots$	$\dots$	$\dots$	$x^{(C_3,N_3)}$
			$\vdots$				
$\prod_{i\in C_p} X_i:$	$x^{(C_p,1)}$	$x^{(C_p,2)}$	$x^{(C_p,3)}$	$\dots$	$x^{(C_p,N_p)}$		

Figure 4.25: Particle families in each $\prod_{i\in C_l} X_i$

In other words, the generated particle family for the product space $\prod_{i=1}^m X_i$ consists of all possible particles in $\prod_{i=1}^m X_i$ otained by choosing the first component from $(x^{(C_1,j_1)})_{j_1=1}^{N_1}$, the second component from $(x^{(C_2,j_2)})_{j_2=1}^{N_2}$, and so on. So a typical particle in the generated particle family could be $(x^{(C_1,3)}, x^{(C_2,8)}, x^{(C_3,42)}, \dots, x^{(C_p,17)})$. There are $\prod_{l=1}^p N_l$ such particles.

Now fix $n \in \mathbb{N}_0$ and $h_n \in H_n$. Suppose that, for $l = 1, \dots, p$, $(x_n^{(C_l,j_l)})_{j_l=1}^{N_l}$ is a particle family in $\prod_{i\in C_l} X_i$ and

$$\nu_n^{(l)}(h_n^{(l)}) = \frac{1}{N_l} \sum_{j_l=1}^{N_l} \delta_{x_n^{(C_l,j_l)}}.$$

Then

$$\begin{aligned}
&\nu_n(h_n) \\
&= \bigotimes_{l=1}^{p} \nu_n^{(l)}(h_n^{(l)}) \\
&= \bigotimes_{l=1}^{p} \frac{1}{N_l} \sum_{j_l=1}^{N_l} \delta_{x_n^{(C_l,j_l)}} \\
&= \frac{1}{\prod_{l=1}^p N_l} \bigotimes_{l=1}^{p} \sum_{j_l=1}^{N_l} \delta_{x_n^{(C_l,j_l)}} \\
&= \frac{1}{\prod_{l=1}^p N_l} \sum_{j_1=1}^{N_1} \cdots \sum_{j_p=1}^{N_p} \delta_{(x_n^{(C_1,j_1)}, x_n^{(C_2,j_2)}, \dots, x_n^{(C_p,j_p)})}.
\end{aligned}$$

Hence

$$\nu_n(h_n) = \frac{1}{\prod_{l=1}^p N_l} \sum_{j_1=1}^{N_1} \cdots \sum_{j_p=1}^{N_p} \delta_{(x_n^{(C_1,j_1)}, x_n^{(C_2,j_2)}, \dots, x_n^{(C_p,j_p)})}.$$

In other words, if

$$\nu_n^{(l)}(h_n^{(l)}) \approx \frac{1}{N_l} \sum_{j_l=1}^{N_l} \delta_{x_n^{(C_l,j_l)}},$$

for $l = 1, \ldots, p$, then

$$\nu_n(h_n) \approx \frac{1}{\prod_{l=1}^{p} N_l} \sum_{j_1=1}^{N_1} \cdots \sum_{j_p=1}^{N_p} \delta_{(x_n^{(C_1,j_1)}, x_n^{(C_2,j_2)}, \ldots, x_n^{(C_p,j_p)})},$$

so that $\nu_n(h_n)$ is approximated by the generated particle family.

In summary so far, under the above assumptions for the factorization of the schema, transition model, and observation model, the distribution $\nu_n(h_n)$ on $\prod_{i=1}^{m} X_i$ is the product of the marginal distributions $\nu_n^{(l)}(h_n^{(l)})$ on the $\prod_{i \in C_l} X_i$. Furthermore, instead of running a particle filter on $\prod_{i=1}^{m} X_i$, one can equivalently run (independent) particle filters on each $\prod_{i \in C_l} X_i$. These are nice results, but the problem is that the conditions under which they hold are too strong for most practical situations. Instead it is necessary to weaken the assumption on the transition model somewhat; the price to be paid is that an approximation is thus introduced.

Consider now the setting as above, except assume instead that, for $l = 1, \ldots, p$, the transition model has the form

$$(\eta_n^{(l)} : H_{n-1}^{(l)} \times A \times \prod_{i=1}^{m} X_i \to \mathcal{P}(\prod_{i \in C_l} X_i))_{n \in \mathbb{N}}.$$

So now the lth transition model depends not just on $\prod_{i \in C_l} X_i$ but on $\prod_{i=1}^{m} X_i$ or, more realistically, on $\prod_{i \in D} X_i$, for some set $D \subseteq \{1, \ldots, m\}$ that is a little larger than C_l. Also

$$\eta_n = \lambda(h, a, x). \bigotimes_{l=1}^{p} \eta_n^{(l)}(h^{(l)}, a, x),$$

for all $n \in \mathbb{N}$. Everything else is the same. This change means that the product of the marginal distributions $\nu_n^{(l)}(h_n^{(l)})$ on the $\prod_{i \in C_l} X_i$ (computed using the modified transition models) is unlikely to be equal to the distribution $\nu_n(h_n)$ on $\prod_{i=1}^{m} X_i$. However, there are applications where the product of the marginal distributions $\nu_n^{(l)}(h_n^{(l)})$ on the $\prod_{i \in C_l} X_i$ is a close approximation of the distribution $\nu_n(h_n)$ on $\prod_{i=1}^{m} X_i$. A factored particle filter exploits this situation.

The factored particle filter is given in Figures 4.26 and 4.27. Figure 4.26 gives the initialization step and Figure 4.27 gives the recursive step of the algorithm. For the initialization, it is assumed that, for $l = 1, \ldots, p$, the initial distribution $\nu_0^{(l)} : H_0 \to \mathcal{P}(\prod_{i \in C_l} X_i)$ is given.

The recursive step in Figure 4.27 is based on the recursive step for the nonconditional filter in Figure 4.17, except now a bank of p particle filters have to be handled, one for each space $\prod_{i \in C_l} X_i$; hence the top-level **do**-loop. The body of this **do**-loop has three parts. The first part samples potential new particles. The main point of interest is the third argument of the function $\eta_n^{(l)}$, which is a random particle from the particle family in $\prod_{i=1}^{m} X_i$ generated by $(x_{n-1}^{(C_1,j_1)})_{j_1=1}^{N_1}, \ldots, (x_{n-1}^{(C_p,j_p)})_{j_p=1}^{N_p}$. Then the weights of these potential new particles are calculated. The second part normalizes the weights, while the third part does the resampling, in essentially the same way as in Figure 4.17.

In the case when the partition is the coarsest one $\{\{1, \ldots, m\}\}$, the algorithm in Figures 4.26 and 4.27 is the same as the algorithm in Figure 4.16 and 4.17. When the

function *InitializeFactoredParticleFilter* **returns** Initial particle families $(x_0^{(C_1,j_1)})_{j_1=1}^{N_1}, \ldots, (x_0^{(C_p,j_p)})_{j_p=1}^{N_p}$;

for $l := 1$ **to** p **do**

  for $j_l := 1$ **to** N_l **do**

    sample $x_0^{(C_l,j_l)} \sim \nu_0^{(l)}(())$;

return $(x_0^{(C_1,j_1)})_{j_1=1}^{N_l}, \ldots, (x_0^{(C_p,j_p)})_{j_p=1}^{N_p}$;

Figure 4.26: Initialization of the factored particle filter for the nonconditional case

function *FactoredParticleFilter*$((x_{n-1}^{(C_1,i_1)})_{i_1=1}^{N_1}, \ldots, (x_{n-1}^{(C_p,i_p)})_{i_p=1}^{N_p}, h_{n-1}, a_n, o_n)$
returns Particle families $(x_n^{(C_1,i_1)})_{i_1=1}^{N_1}, \ldots, (x_n^{(C_p,i_p)})_{i_p=1}^{N_p}$ at time n;

inputs: Particle families $(x_{n-1}^{(C_1,i_1)})_{i_1=1}^{N_1}, \ldots, (x_{n-1}^{(C_p,i_p)})_{i_p=1}^{N_p}$ at time $n-1$,
history h_{n-1} up to time $n-1$,
action a_n at time n,
observation o_n at time n;

for $l := 1$ **to** p **do**

  for $i_l := 1$ **to** N_l **do**

    sample $\overline{x}_n^{(C_l,i_l)} \sim \frac{1}{\prod_{l=1}^{p} N_l} \sum_{i'_1=1}^{N_1} \cdots \sum_{i'_p=1}^{N_p} \eta_n^{(l)}(h_{n-1}^{(l)}, a_n, (x_{n-1}^{(C_1,i'_1)}, \ldots, x_{n-1}^{(C_p,i'_p)}))$;
    $\widetilde{w}_n^{(C_l,i_l)} := \breve{\zeta}_n^{(l)}(h_{n-1}^{(l)}, a_n, \overline{x}_n^{(C_l,i_l)})(o_n^{(l)})$;

  for $i_l := 1$ **to** N_l **do**

    $w_n^{(C_l,i_l)} := \frac{\widetilde{w}_n^{(C_l,i_l)}}{\sum_{i'_l=1}^{N_l} \widetilde{w}_n^{(C_l,i'_l)}}$;

  for $i_l := 1$ **to** N_l **do**

    sample $x_n^{(C_l,i_l)} \sim \sum_{i'_l=1}^{N_l} w_n^{(C_l,i'_l)} \delta_{\overline{x}_n^{(C_l,i'_l)}}$;

return $(x_n^{(C_1,i_1)})_{i_1=1}^{N_1}, \ldots, (x_n^{(C_p,i_p)})_{i_p=1}^{N_p}$;

Figure 4.27: Recursive step of the factored particle filter for the nonconditional case

partition is the finest one $\{\{1\},\ldots,\{m\}\}$, the algorithm in Figures 4.26 and 4.27 is called the *fully factored particle filter*. The factored particle filter based on the partition $\mathfrak{A}$ of $\{1,\ldots,m\}$ is called the $\mathfrak{A}$-*factored particle filter*

A question relevant to Part 1 of Proposition 4.5.1 is to ask how close the schema on the entire product space is to being exactly factorizable if the assumption that

$$\mathsf{P}(\bigcap_{l=1}^{p} \mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid \mathbf{h}_n) = \prod_{l=1}^{p} \mathsf{P}(\mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid \mathbf{h}_n^{(l)}) \text{ a.s.},$$

for all $A_l \in \bigotimes_{i\in C_l} \mathcal{A}_i$ and for $l = 1,\ldots,p$, holds only approximately. This question has a simple answer.

Towards this, fix $A_l \in \bigotimes_{i\in C_l} \mathcal{A}_i$, for $l = 1,\ldots,p$. Suppose that there exists $\epsilon > 0$ such that, almost surely,

$$\left| \mathsf{P}(\bigcap_{l=1}^{p} \mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid \mathbf{h}_n) - \prod_{l=1}^{p} \mathsf{P}(\mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid \mathbf{h}_n^{(l)}) \right| < \epsilon.$$

Thus the conditional independence assumption is assumed to hold only approximately. Let $(\nu_n : H_n \to \mathcal{P}(\prod_{i=1}^{m} X_i))_{n\in\mathbb{N}_0}$ be the schema for $(\mathbf{x}^{(C_1)},\ldots,\mathbf{x}^{(C_p)})$. Thus

$$\mathsf{P}(\bigcap_{l=1}^{p} \mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid \mathbf{h}_n) = \lambda\omega.\nu_n(\mathbf{h}_n(\omega))(\prod_{l=1}^{p}(A_l) \text{ a.s.}$$

Also, for $l = 1,\ldots,p$, since $(\nu_n^{(l)} : H_n^{(l)} \to \mathcal{P}(\prod_{i\in C_l} X_i))_{n\in\mathbb{N}_0}$ is the schema for $\mathbf{x}^{(C_l)}$, it follows that

$$\mathsf{P}(\mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid \mathbf{h}_n^{(l)}) = \lambda\omega.\nu_n^{(l)}(\mathbf{h}_n^{(l)}(\omega))(A_l) \text{ a.s.}$$

Then, almost surely,

$$\begin{aligned} &\left| \lambda\omega.\nu_n(\mathbf{h}_n(\omega))(\prod_{l=1}^{p} A_l) - \prod_{l=1}^{p} \lambda\omega.\nu_n^{(l)}(\mathbf{h}_n^{(l)}(\omega))(A_l) \right| \\ = &\left| \mathsf{P}(\bigcap_{l=1}^{p} \mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid \mathbf{h}_n) - \prod_{l=1}^{p} \mathsf{P}(\mathbf{x}_n^{(C_l)^{-1}}(A_l) \mid \mathbf{h}_n^{(l)}) \right| \\ < &\epsilon. \end{aligned}$$

This provides an estimate for how close ν_n is to having an exact factorization in terms of the $\nu_n^{(l)}$. Clearly, there is a similar analysis for transition and observation models.

4.6 Factored Conditional Particle Filters

Attention now turns to factored particle filters in the conditional case. The setting for this section is where the schema has the form

$$(\mu_n : H_n \times X \to \mathcal{P}(\prod_{i=1}^{m} Y_i))_{n\in\mathbb{N}_0},$$

the transition model has the form

$$(\tau_n : H_{n-1} \times A \times X \times \prod_{i=1}^{m} Y_i \to \mathcal{P}(\prod_{i=1}^{m} Y_i))_{n\in\mathbb{N}},$$

and the observation model has the form

$$(\xi_n : H_{n-1} \times A \times X \times \prod_{i=1}^{m} Y_i \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

The main interest is when $\prod_{i=1}^{m} Y_i$ is a high-dimensional space.

In addition, there is a schema having the form

$$(\nu_n : H_n \to \mathcal{P}(X)_{n\in\mathbb{N}_0},$$

a transition model having the form

$$(\eta_n : H_{n-1} \times A \times X \to \mathcal{P}(X))_{n\in\mathbb{N}},$$

and an observation model having the form

$$(\zeta_n : H_{n-1} \times A \times X \to \mathcal{P}(O))_{n\in\mathbb{N}}.$$

Typically, X is a parameter space.

The schema μ, the transition model τ, and the observation model ξ are assumed to have particular factorizable forms depending on the choice of partition $\{C_1, \ldots, C_p\}$. For $l = 1, \ldots, p$, let

$$(\mu_n^{(l)} : H_n^{(l)} \times X \to \mathcal{P}(\prod_{i\in C_l} Y_i))_{n\in\mathbb{N}_0},$$

be a schema for which

$$\mu_n = \lambda(h, x).\bigotimes_{l=1}^{p} \mu_n^{(l)}(h^{(l)}, x),$$

for all $n \in \mathbb{N}_0$. Also, for $l = 1, \ldots, p$, suppose that

$$(\tau_n^{(l)} : H_{n-1}^{(l)} \times A \times X \times \prod_{i\in C_l} Y_i \to \mathcal{P}(\prod_{i\in C_l} Y_i))_{n\in\mathbb{N}},$$

is a transition model for which

$$\tau_n = \lambda(h, a, x, (y_{C_1}, \ldots, y_{C_l})).\bigotimes_{l=1}^{p} \tau_n^{(l)}(h^{(l)}, a, x, y_{C_l}),$$

for all $n \in \mathbb{N}$. Similarly, for $l = 1, \ldots, p$, suppose that

$$(\xi_n^{(l)} : H_{n-1}^{(l)} \times A \times X \times \prod_{i\in C_l} Y_i \to \mathcal{P}(O_l))_{n\in\mathbb{N}},$$

is an observation model for which

$$\xi_n = \lambda(h, a, x, (y_{C_1}, \ldots, y_{C_l})).\bigotimes_{l=1}^{p} \xi_n^{(l)}(h^{(l)}, a, x, y_{C_l}),$$

for all $n \in \mathbb{N}$. Thus there is a schema $\mu^{(l)}$, transition model $\tau^{(l)}$, and observation model $\xi^{(l)}$ for each $\prod_{i\in C_l} Y_i$ that is used to define the schema μ, the transition model τ, and the observation model ξ, respectively, for the entire product space $\prod_{i=1}^{m} Y_i$.

The above setting is actually possible, that is, under certain circumstances, there do exist schemas, transition models, and observation models that satisfy the above factorization conditions, as Proposition 4.6.1 below shows. In each case of schema, transition model, and observation model, there is a conditional independence assumption that is sufficient to ensure the desired factorization holds.

Proposition 4.6.1. *(Factorization for the conditional case) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $\{C_1, \ldots, C_p\}$ a partition of $\{1, \ldots, m\}$, $(A, \mathcal{A})$ an action space, $(O_l, \mathcal{O}_l)$ an observation space, for $l = 1, \ldots, p$, $(X_i, \mathcal{A}_i)$ and $(Y_i, \mathcal{B}_i)$ measurable spaces, for $i = 1, \ldots, m$, $\mathbf{a} : \Omega \to A^{\mathbb{N}}$ an action process, $\mathbf{x} : \Omega \to X^{\mathbb{N}}$ a stochastic process, and $\mathbf{o}^{(l)} : \Omega \to O_l^{\mathbb{N}}$ an observation process and $\mathbf{y}^{(C_l)} : \Omega \to (\prod_{i\in C_l} Y_i)^{\mathbb{N}_0}$ a stochastic process, for $l = 1, \ldots, p$.*

1. Let $(\mu_n^{(l)} : H_n^{(l)} \times X \to \mathcal{P}(\prod_{i\in C_l} Y_i))_{n\in\mathbb{N}_0}$ be the schema for $\mathbf{y}^{(C_l)}$ given $\mathbf{x}$, for $l = 1, \ldots, p$, and

$$\mu_n \triangleq \lambda(h, x).\bigotimes_{l=1}^{p} \mu_n^{(l)}(h^{(l)}, x) : H_n \times X \to \mathcal{P}(\prod_{i=1}^{m} Y_i),$$

for all $n \in \mathbb{N}_0$. Suppose that

$$\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid (\mathbf{h}_n, \mathbf{x}_n)) = \prod_{l=1}^{p} \mathsf{P}({\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid (\mathbf{h}_n^{(l)}, \mathbf{x}_n)) \text{ a.s.},$$

for all $B_l \in \bigotimes_{i\in C_l} \mathcal{B}_i$ and for $l = 1, \ldots, p$. Then $(\mu_n : H_n \times X \to \mathcal{P}(\prod_{i=1}^{m} Y_i))_{n\in\mathbb{N}_0}$ is the schema for $(\mathbf{y}^{(C_1)}, \ldots, \mathbf{y}^{(C_p)})$ given $\mathbf{x}$.

2. Let $(\tau_n^{(l)} : H_{n-1}^{(l)} \times A \times X \times \prod_{i\in C_l} Y_i \to \mathcal{P}(\prod_{i\in C_l} Y_i))_{n\in\mathbb{N}}$ be the transition model for $\mathbf{y}^{(C_l)}$ given $\mathbf{x}$, for $l = 1, \ldots, p$, and

$$\tau_n \triangleq \lambda(h, a, x, (y_{C_1}, \ldots, y_{C_p})).\bigotimes_{l=1}^{p} \tau_n^{(l)}(h^{(l)}, a, x, y_{C_l}) : H_{n-1} \times A \times X \times \prod_{i=1}^{m} Y_i \to \mathcal{P}(\prod_{i=1}^{m} Y_i),$$

for all $n \in \mathbb{N}$. Suppose that

$$\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_{n-1}^{(C_1)}, \ldots, \mathbf{y}_{n-1}^{(C_p)}))) =$$

$$\prod_{l=1}^{p} \mathsf{P}({\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_l)})) \text{ a.s.},$$

for all $B_l \in \bigotimes_{i\in C_l} \mathcal{B}_i$ *and for* $l = 1, \ldots, p$. *Then* $(\tau_n : H_{n-1} \times A \times X \times \prod_{i=1}^m Y_i \to \mathcal{P}(\prod_{i=1}^m X_i))_{n\in\mathbb{N}}$ *is the transition model for* $(\mathbf{y}^{(C_1)}, \ldots, \mathbf{y}^{(C_p)})$ *given* $\mathbf{x}$.

3. Let $(\xi_n^{(l)} : H_{n-1}^{(l)} \times A \times X \times \prod_{i\in C_l} Y_i \to \mathcal{P}(O_l))_{n\in\mathbb{N}}$ *be the observation model for* $\mathbf{y}^{(C_l)}$ *given* $\mathbf{x}$, *for* $l = 1, \ldots, p$, *and*

$$\xi_n \triangleq \lambda(h, a, x, (y_{C_1}, \ldots, y_{C_p})).\bigotimes_{l=1}^{p} \zeta_n^{(l)}(h^{(l)}, a, x, y_{C_l}) : H_{n-1} \times A \times X \times \prod_{i=1}^{m} Y_i \to \mathcal{P}(\prod_{l=1}^{p} O_l),$$

for all $n \in \mathbb{N}$. *Suppose that*

$$\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{o}_n^{(l)}}^{-1}(D_l) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_n^{(C_1)}, \ldots, \mathbf{y}_n^{(C_p)}))) = \prod_{l=1}^{p} \mathsf{P}({\mathbf{o}_n^{(l)}}^{-1}(D_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_l)}))\ a.s.,$$

for all $D_l \in \mathcal{O}_l$ *and for* $l = 1, \ldots, p$. *Then* $(\xi_n : H_{n-1} \times A \times X \times \prod_{i=1}^m Y_i \to \mathcal{P}(\prod_{l=1}^p O_l))_{n\in\mathbb{N}}$ *is the observation model for* $(\mathbf{y}^{(C_1)}, \ldots, \mathbf{y}^{(C_p)})$ *given* $\mathbf{x}$.

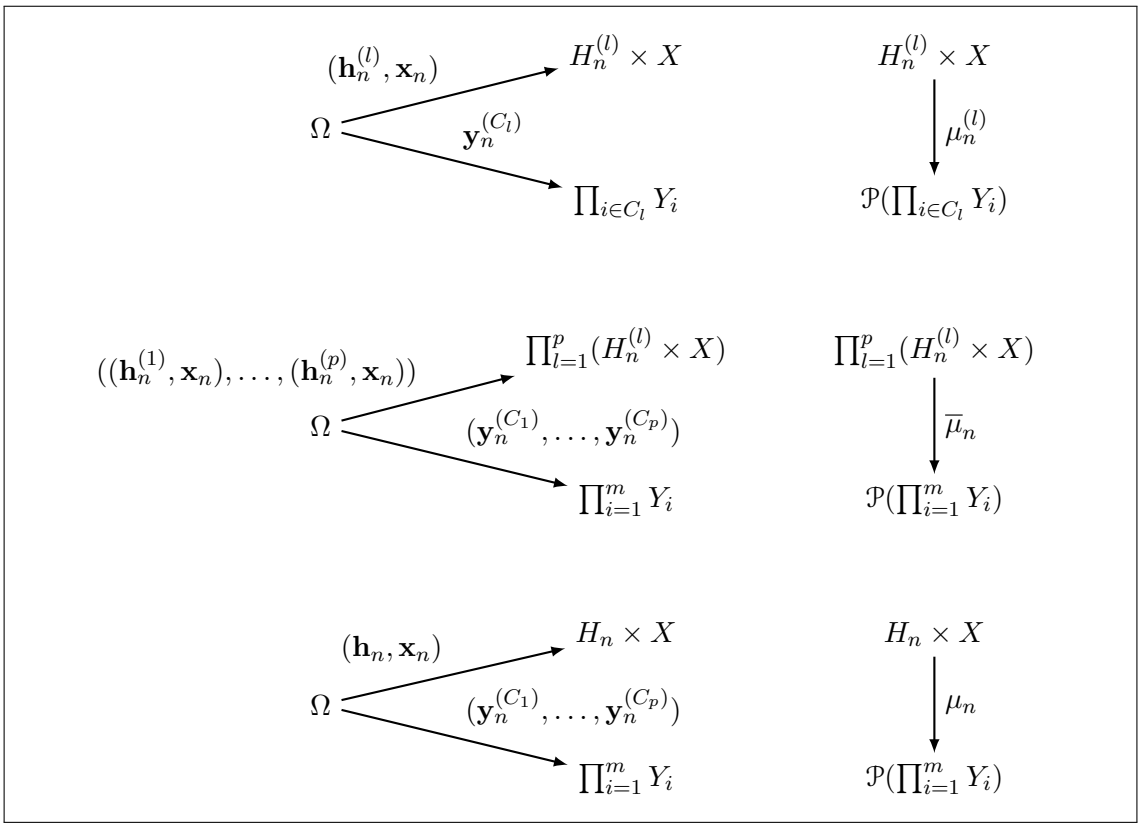

Figure 4.28: Setting for Part 1 of Proposition 4.6.1

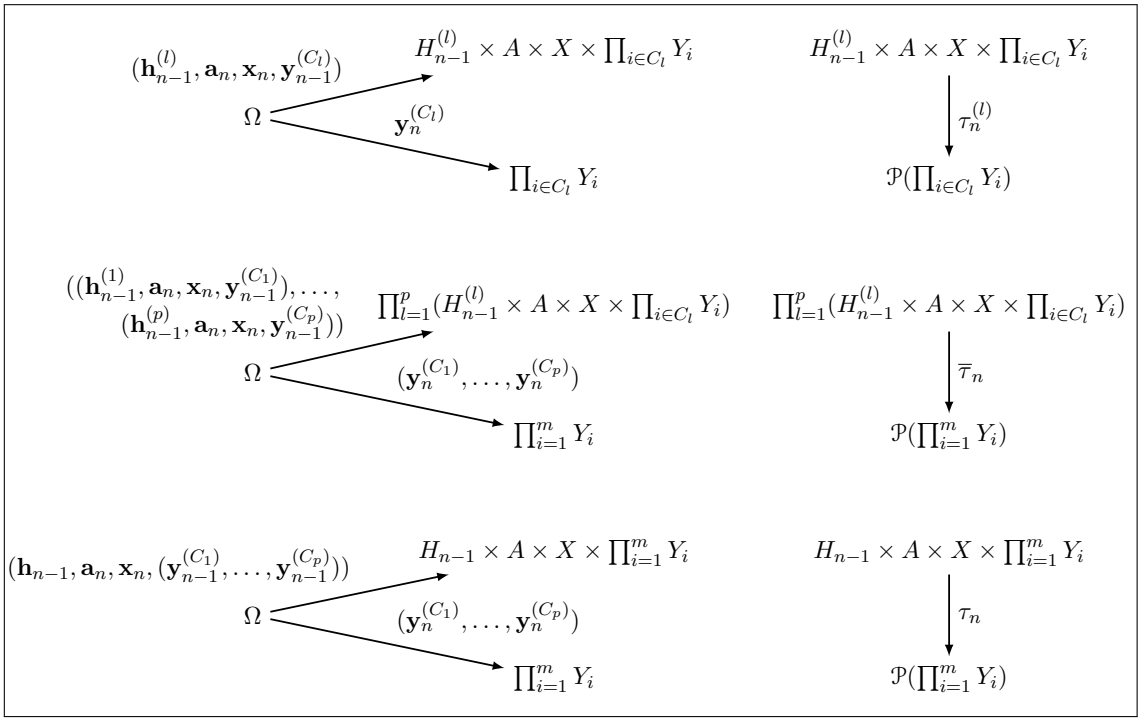

Figure 4.29: Setting for Part 2 of Proposition 4.6.1

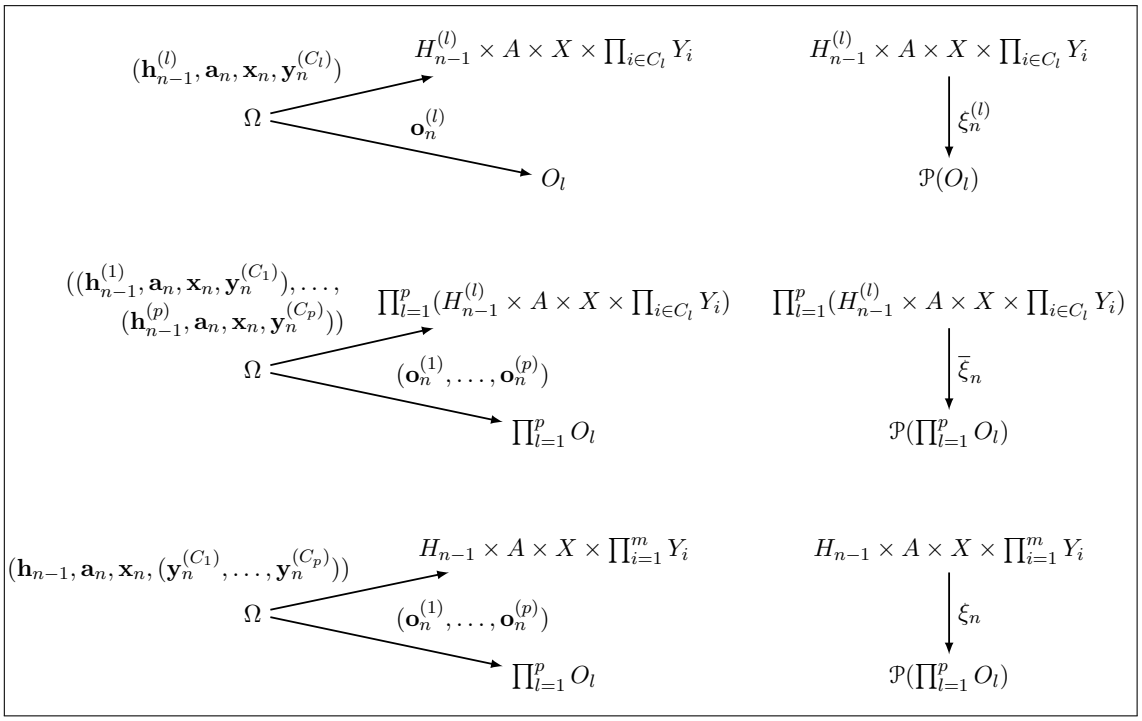

Figure 4.30: Setting for Part 3 of Proposition 4.6.1

Proof. 1. Since $(\mu_n^{(l)} : H_n^{(l)} \times X \to \mathcal{P}(\prod_{i\in C_l} Y_i))_{n\in\mathbb{N}_0}$ is the schema for $\mathbf{y}^{(C_l)}$ given $\mathbf{x}$, it follows that

$$\mathsf{P}({\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid (\mathbf{h}_n^{(l)}, \mathbf{x}_n)) = \lambda\omega.\mu_n^{(l)}((\mathbf{h}_n^{(l)}, \mathbf{x}_n)(\omega))(B_l) \text{ a.s.},$$

for all $B_l \in \bigotimes_{i\in C_l} \mathcal{B}_i$ and for $l = 1, \ldots, p$. Also, since

$$\sigma((\mathbf{h}_n, \mathbf{x}_n)) = \sigma(((\mathbf{h}_n^{(1)}, \mathbf{x}_n), \ldots, (\mathbf{h}_n^{(p)}, \mathbf{x}_n))),$$

it follows that

$$\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid ((\mathbf{h}_n^{(1)}, \mathbf{x}_n), \ldots, (\mathbf{h}_n^{(p)}, \mathbf{x}_n))) = \prod_{l=1}^{p} \mathsf{P}({\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid (\mathbf{h}_n^{(l)}, \mathbf{x}_n)) \text{ a.s.},$$

for all $B_l \in \bigotimes_{i\in C_l} \mathcal{B}_i$ and for $l = 1, \ldots, p$.

Let

$$\overline{\mu}_n \triangleq \lambda((h_1, x_1), \ldots, (h_p, x_p)).\bigotimes_{l=1}^{p} \mu_n^{(l)}(h_l, x_l) : \prod_{l=1}^{p}(H_n^{(l)} \times X) \to \mathcal{P}(\prod_{i=1}^{m} Y_i).$$

Then, for all $B \in \bigotimes_{i=1}^{m} \mathcal{B}_i$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}((\mathbf{y}_n^{(C_1)}, \ldots, \mathbf{y}_n^{(C_p)})^{-1}(B) \mid (\mathbf{h}_n, \mathbf{x}_n)) \\
&= \mathsf{P}((\mathbf{x}_n^{(C_1)}, \ldots, \mathbf{x}_n^{(C_p)})^{-1}(B) \mid ((\mathbf{h}_n^{(1)}, \mathbf{x}_n), \ldots, (\mathbf{h}_n^{(p)}, \mathbf{x}_n))) \\
&\qquad\qquad [\sigma((\mathbf{h}_n, \mathbf{x}_n)) = \sigma(((\mathbf{h}_n^{(1)}, \mathbf{x}_n), \ldots, (\mathbf{h}_n^{(p)}\mathbf{x}_n)))] \\
&= \lambda\omega.\overline{\mu}_n(((\mathbf{h}_n^{(1)}, \mathbf{x}_n), \ldots, (\mathbf{h}_n^{(p)}, \mathbf{x}_n))(\omega))(B) \qquad \text{[Proposition A.5.22]} \\
&= \lambda\omega.\bigotimes_{l=1}^{p} \mu_n^{(l)}((\mathbf{h}_n^{(l)}, \mathbf{x}_n)(\omega))(B) \\
&= \lambda\omega.\mu_n((\mathbf{h}_n, \mathbf{x}_n)(\omega))(B).
\end{aligned}$$

That is, $(\mu_n : H_n \times X \to \mathcal{P}(\prod_{i=1}^{m} Y_i))_{n\in\mathbb{N}_0}$ is the schema for $(\mathbf{y}^{(C_1)}, \ldots, \mathbf{x}^{(C_p)})$ given $\mathbf{x}$.

2. Since $(\tau_n^{(l)} : H_{n-1}^{(l)} \times A \times X \times \prod_{i\in C_l} Y_i \to \mathcal{P}(\prod_{i\in C_l} Y_i))_{n\in\mathbb{N}}$ is the transition model for $\mathbf{y}^{(C_l)}$ given $\mathbf{x}$, it follows that

$$\mathsf{P}({\mathbf{y}_n^{(C_l)}}^{-1}(A_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_l)})) = \lambda\omega.\tau_n^{(l)}((\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_l)})(\omega))(B_l) \text{ a.s.},$$

for all $B_l \in \bigotimes_{i\in C_l} \mathcal{B}_i$ and for $l = 1, \ldots, p$. Also, since

$$\sigma((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_{n-1}^{(C_1)}, \ldots, \mathbf{y}_{n-1}^{(C_p)}))) = \sigma(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_p)}))),$$

it follows that

$$\begin{aligned}
&\mathsf{P}(\bigcap_{l=1}^{p} {\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid ((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_p)}))) = \\
&\qquad\qquad \prod_{l=1}^{p} \mathsf{P}({\mathbf{y}_n^{(C_l)}}^{-1}(B_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_l)})) \text{ a.s.},
\end{aligned}$$

for all $B_l \in \bigotimes_{i\in C_l} \mathcal{B}_i$ and for $l = 1, \ldots, p$.

Let

$$\overline{\tau}_n \triangleq \lambda((h_1, a_1, x_1, y_1), \ldots, (h_p, a_p, x_p, y_p)).\bigotimes_{l=1}^{p} \tau_n^{(l)}(h_l, a_l, x_l, y_l) : \prod_{l=1}^{p}(H_{n-1}^{(l)} \times A \times X \times \prod_{i\in C_l} Y_i) \to \mathcal{P}(\prod_{i=1}^{m} Y_i).$$

Then, for all $B \in \bigotimes_{i=1}^{m} \mathcal{B}_i$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}((\mathbf{y}_n^{(C_1)}, \ldots, \mathbf{y}_n^{(C_p)})^{-1}(B) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_{n-1}^{(C_1)}, \ldots, \mathbf{y}_{n-1}^{(C_p)})))\\
&= \mathsf{P}((\mathbf{y}_n^{(C_1)}, \ldots, \mathbf{y}_n^{(C_p)})^{-1}(B) \mid ((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_p)})))\\
&\qquad [\sigma((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_{n-1}^{(C_1)}, \ldots, \mathbf{y}_{n-1}^{(C_p)}))) =\\
&\qquad\qquad \sigma(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_p)})))]\\
&= \lambda\omega.\overline{\tau}_n(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_p)}))(\omega))(B)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[Proposition A.5.22]}\\
&= \lambda\omega.\bigotimes_{l=1}^{p} \tau_n^{(l)}((\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_{n-1}^{(C_l)})(\omega))(B)\\
&= \lambda\omega.\tau_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_{n-1}^{(C_1)}, \ldots, \mathbf{y}_{n-1}^{(C_p)}))(\omega))(B).
\end{aligned}$$

That is, $(\tau_n : H_{n-1} \times A \times X \times \prod_{i=1}^{m} Y_i \to \mathcal{P}(\prod_{i=1}^{m} Y_i))_{n\in\mathbb{N}}$ is the transition model for $(\mathbf{y}^{(C_1)}, \ldots, \mathbf{y}^{(C_p)})$ given $\mathbf{x}$.

3. Since $(\xi_n^{(l)} : H_{n-1}^{(l)} \times A \times X \times \prod_{i\in C_l} Y_i \to \mathcal{P}(O_l)_{n\in\mathbb{N}}$ is the observation model for $\mathbf{y}^{(C_l)}$ given $\mathbf{x}$, it follows that

$$\mathsf{P}(\mathbf{o}_n^{(l)^{-1}}(D_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_l)})) = \lambda\omega.\xi_n^{(l)}((\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_l)})(\omega))(D_l) \text{ a.s.},$$

for all $D_l \in \mathcal{O}_l$ and for $l = 1, \ldots, p$. Also, since

$$\sigma((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_n^{(C_1)}, \ldots, \mathbf{y}_n^{(C_p)}))) = \sigma(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_p)}))),$$

it follows that

$$\mathsf{P}(\bigcap_{l=1}^{p} \mathbf{o}_n^{(l)^{-1}}(D_l) \mid ((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_p)}))) = \prod_{l=1}^{p} \mathsf{P}(\mathbf{o}_n^{(l)^{-1}}(D_l) \mid (\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_l)})) \text{ a.s.},$$

for all $D_l \in \mathcal{O}_l$ and for $l = 1, \ldots, p$.

Let

$$\overline{\xi}_n \triangleq \lambda((h_1, a_1, x_1, y_1), \ldots, (h_p, a_p, x_p, y_p)).\bigotimes_{l=1}^{p} \xi_n^{(l)}(h_l, a_l, x_l, y_l) :$$
$$\prod_{l=1}^{p}(H_{n-1}^{(l)} \times A \times X \times \prod_{i \in C_l} Y_i) \to \mathcal{P}(\prod_{l=1}^{p} O_l).$$

Then, for all $D \in \bigotimes_{l=1}^{p} \mathcal{O}_l$, P-almost surely,

$$\begin{aligned}
&\mathsf{P}((\mathbf{o}_n^{(1)}, \ldots, \mathbf{o}_n^{(p)})^{-1}(D) \mid (\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_n^{(C_1)}, \ldots, \mathbf{y}_n^{(C_p)}))) \\
&= \mathsf{P}((\mathbf{o}_n^{(1)}, \ldots, \mathbf{o}_n^{(p)})^{-1}(D) \mid ((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_p)}))) \\
&\qquad [\sigma((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_n^{(C_1)}, \ldots, \mathbf{y}_n^{(C_p)}))) = \\
&\qquad\qquad \sigma(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_p)})))] \\
&= \lambda\omega.\overline{\xi}_n(((\mathbf{h}_{n-1}^{(1)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_1)}), \ldots, (\mathbf{h}_{n-1}^{(p)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_p)}))(\omega))(D) \\
&\qquad\qquad [\text{Proposition A.5.22}] \\
&= \lambda\omega.\bigotimes_{l=1}^{p} \xi_n^{(l)}((\mathbf{h}_{n-1}^{(l)}, \mathbf{a}_n, \mathbf{x}_n, \mathbf{y}_n^{(C_l)})(\omega))(D) \\
&= \lambda\omega.\xi_n((\mathbf{h}_{n-1}, \mathbf{a}_n, \mathbf{x}_n, (\mathbf{y}_n^{(C_1)}, \ldots, \mathbf{y}_n^{(C_p)}))(\omega))(D).
\end{aligned}$$

That is, $(\xi_n : H_{n-1} \times A \times X \times \prod_{i=1}^{m} Y_i \to \mathcal{P}(\prod_{l=1}^{p} O_l))_{n \in \mathbb{N}}$ is the observation model for $(\mathbf{y}^{(C_1)}, \ldots, \mathbf{y}^{(C_p)})$ given $\mathbf{x}$. □

In summary so far, under the above assumptions for the factorization of the schema, transition model, and observation model, the distribution $\mu_n(h_n, x)$ on $\prod_{i=1}^{m} Y_i$ is the product of the marginal distributions $\mu_n^{(l)}(h_n^{(l)}, x)$ on the $\prod_{i \in C_l} Y_i$. Furthermore, instead of running a conditional particle filter on $\prod_{i=1}^{m} Y_i$, one can equivalently run (independent) conditional particle filters on each $\prod_{i \in C_l} Y_i$. Once again, as for the nonconditional case, the assumption on the transition model will be weakened; again this introduces an approximation.

Consider now the setting as above, except assume instead that, for $l = 1, \ldots, p$, the transition model has the form

$$(\tau_n^{(l)} : H_{n-1}^{(l)} \times A \times X \times \prod_{i=1}^{m} Y_i \to \mathcal{P}(\prod_{i \in C_l} Y_i))_{n \in \mathbb{N}}.$$

So now the lth transition model depends not just on $\prod_{i \in C_l} Y_i$ but on $\prod_{i=1}^{m} Y_i$. Also

$$\tau_n = \lambda(h, a, x, y).\bigotimes_{l=1}^{p} \tau_n^{(l)}(h^{(l)}, a, x, y),$$

for all $n \in \mathbb{N}$. Everything else is the same. Once again, this change will mean that the product of the marginal distributions $\mu_n^{(l)}(h_n^{(l)}, x)$ on the $\prod_{i \in C_l} Y_i$ (computed using the

modified transition model) is unlikely to be equal to the distribution $\mu_n(h_n, x)$ on $\prod_{i=1}^m X_i$. However, there are applications where the product of the marginal distributions $\mu_n^{(l)}(h_n^{(l)}, x)$ on the $\prod_{i\in C_l} Y_i$ is a close approximation of the distribution $\mu_n(h_n, x)$ on $\prod_{i=1}^m Y_i$. The conditional factored particle filter exploits this situation.

Conditional particles and conditional particle families that are factored will be needed. In the following three definitions, let $\{C_1, \ldots, C_p\}$ be a partition of the index set $\{1, \ldots, m\}$ for the product space $\prod_{i=1}^m Y_i$.

Definition 4.6.1. A *factored particle family* (with respect to $\{C_1, \ldots, C_p\}$ in $\prod_{i=1}^m Y_i$) is a p-tuple

$$((y^{(C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y^{(C_p,j_p)})_{j_p=1}^{M_p}),$$

where $(y^{(C_l,j_l)})_{j_l=1}^{M_l}$ is a particle family in $\prod_{i\in C_l} Y_i$, for $l = 1, \ldots, p$.

Definition 4.6.2. A *factored conditional particle* (with respect to $\{C_1, \ldots, C_p\}$ from X to $\prod_{i=1}^m Y_i$) is a pair

$$(x, ((y^{(C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y^{(C_p,j_p)})_{j_p=1}^{M_p})),$$

where x is a particle in X and $((y^{(C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y^{(C_p,j_p)})_{j_p=1}^{M_p}))$ is a factored particle family in $\prod_{i=1}^m Y_i$.

Definition 4.6.3. A *factored conditional particle family* (with respect to $\{C_1, \ldots, C_p\}$ from X to $\prod_{i=1}^m Y_i$) is an indexed family of the form

$$(x^{(k)}, ((y^{(k,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y^{(k,C_p,j_p)})_{j_p=1}^{M_p}))_{k=1}^N,$$

where $(x^{(k)}, ((y^{(k,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y^{(k,C_p,j_p)})_{j_p=1}^{M_p}))$ is a factored conditional particle, for $k = 1, \ldots, N$.

So a factored conditional particle family is an indexed family of factored conditional particles.

It is assumed that, for $l = 1, \ldots, p$, the initial conditional distributions $\mu_0^{(l)} : H_0 \times X \to \mathcal{P}(\prod_{i\in C_l} Y_i)$ are given. Figure 4.31 gives the initialization step for the factored conditional particle filter.

Figure 4.32 gives the recursive step for the factored conditional particle filter. The recursive step in Figure 4.32 is based on the recursive step for the nonconditional filter in Figure 4.21, except now a bank of p particle filters have to be handled, one for each space $\prod_{i\in C_l} Y_i$; hence the second-level **do**-loop. The body of this second-level **do**-loop has three parts. The first part samples potential new particles. The main point of interest is the fourth argument of the function $\tau_n^{(l)}$, which is a random particle from the particle family in $\prod_{i=1}^m Y_i$ generated by $(y_{n-1}^{(k^*,C_1,j_1)})_{j_1=1}^{N_1}, \ldots, (y_{n-1}^{(k^*,C_p,j_p)})_{j_p=1}^{N_p}$. Then the weights of these potential new particles are calculated. The second part normalizes the weights, while the third part does the resampling, in essentially the same way as in Figure 4.21.

In the case when the partition is the coarsest one $\{\{1, \ldots, m\}\}$, the algorithm in Figures 4.31 and 4.32 is the same as the algorithm in Figure 4.20 and 4.21 . When the

function *InitializeFactoredConditionalParticleFilter*$((x_0^{(k)})_{k=1}^N)$
returns Initial factored conditional particle family
$(x_0^{(k)}, ((y_0^{(k,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y_0^{(k,C_p,j_p)})_{j_p=1}^{M_p}))_{i=1}^N$;
input: Initial particle family $(x_0^{(k)})_{k=1}^N$;

for $k := 1$ **to** N **do**
  for $l := 1$ **to** p **do**
    for $j_l := 1$ **to** M_l **do**
      sample $y_0^{(k,C_l,j_l)} \sim \mu_0^{(l)}((), x_0^{(k)})$;
return $(x_0^{(k)}, ((y_0^{(k,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y_0^{(k,C_p,j_p)})_{j_p=1}^{M_p}))_{k=1}^N$;

Figure 4.31: Initialization of the factored particle filter for the conditional case

partition is the finest one $\{\{1\}, \ldots, \{m\}\}$, the algorithm in Figures 4.31 and 4.32 is called the *fully factored conditional particle filter*. The factored conditional particle filter based on the partition $\mathfrak{A}$ of $\{1, \ldots, m\}$ is called the $\mathfrak{A}$-*factored conditional particle filter*

Finally in this section, the issue of learning parameters using a factored particle filter is discussed. For the parameter space X, assuming a jittering transition model is used, the appropriate algorithm is given in Figures 4.16 and 4.19. For the state space $Y \triangleq \prod_{i=1}^m Y_i$, the appropriate algorithm is given in Figures 4.31 and 4.32.

The discussion towards the end of Section 4.3 about using the observation model synthesis proposition to handle the assignment statement

$$\widetilde{w}_n^{(i)} := \breve{\zeta}_n(h_{n-1}, a_n, \overline{x}_n^{(i)})(o_n);$$

in Figure 4.19 is just as relevant here. The only difference is that instead of approximating using a conditional particle

$$(x_{n-1}^{(i^*)}, (y_{n-1}^{(i^*,j)})_{j=1}^M),$$

one uses instead a factored conditional particle

$$(x_{n-1}^{(i^*)}, ((y_{n-1}^{(i^*,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y_{n-1}^{(i^*,C_p,j_p)})_{j_p=1}^{M_p})).$$

Whenever an approximation of the state distribution corresponding to the parameter $x_{n-1}^{(i^*)}$ is needed, one samples sufficient particles from the particle family generated by

$$((y_{n-1}^{(i^*,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y_{n-1}^{(i^*,C_p,j_p)})_{j_p=1}^{M_p}).$$

To do: Incorporate the 12 filtering algorithms in [25] into this chapter.

Bibliographical Notes

Belief acquisition as envisaged in this book is a generalization of filtering that was originally invented in the field of signal processing [84]. For an introductory account of filtering, see

function *FactoredConditionalParticleFilter*
$((x_n^{(i)})_{i=1}^N, (x_{n-1}^{(i)}, ((y_{n-1}^{(i,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y_{n-1}^{(i,C_p,j_p)})_{j_p=1}^{M_p}))_{i=1}^N, h_{n-1}, a_n, o_n)$
returns Factored conditional particle family
$(x_n^{(i)}, ((y_n^{(i,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y_n^{(i,C_p,j_p)})_{j_p=1}^{M_p}))_{i=1}^N$ at time n;

inputs: Particle family $(x_n^{(i)})_{i=1}^N$ at time n,
factored conditional particle family
$(x_{n-1}^{(i)}, ((y_{n-1}^{(i,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y_{n-1}^{(i,C_p,j_p)})_{j_p=1}^{M_p}))_{i=1}^N$ at time $n-1$,
history h_{n-1} up to time $n-1$,
action a_n at time n,
observation o_n at time n;

for $i := 1$ **to** N **do**
  for $l := 1$ **to** p **do**
    for $j_l := 1$ **to** M_l **do**
      sample $\overline{y}_n^{(i,C_l,j_l)} \sim$
      $\frac{1}{\prod_{l=1}^p M_l} \sum_{j'_1=1}^{M_1} \cdots \sum_{j'_p=1}^{M_p} \tau_n^{(l)}(h_{n-1}^{(l)}, a_n, x_n^{(i)}, (y_{n-1}^{(i^*,C_1,j'_1)}, \ldots, y_{n-1}^{(i^*,C_p,j'_p)}))$;

      $\widetilde{w}_n^{(i,C_l,j_l)} := \breve{\xi}_n^{(l)}(h_{n-1}^{(l)}, a_n, x_n^{(i)}, \overline{y}_n^{(i,C_l,j_l)})(o_n^{(l)})$;
    for $j_l := 1$ **to** M_l **do**
      $w_n^{(i,C_l,j_l)} := \frac{\widetilde{w}_n^{(i,C_l,j_l)}}{\sum_{j'_l=1}^{N_l} \widetilde{w}_n^{(i,C_l,j'_l)}}$;
    for $j_l := 1$ **to** M_l **do**
      sample $y_n^{(i,C_l,j_l)} \sim \sum_{j'_l=1}^{M_l} w_n^{(i,C_l,j'_l)} \delta_{\overline{y}_n^{(i,C_l,j'_l)}}$;

return $(x_n^{(i)}, ((y_n^{(i,C_1,j_1)})_{j_1=1}^{M_1}, \ldots, (y_n^{(i,C_p,j_p)})_{j_p=1}^{M_p}))_{i=1}^N$

Figure 4.32: Recursive step of the factored particle filter for the conditional case, where $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s.

[141]; in particular, the relationship between filtering and Bayesian inference is discussed there. Filtering has been heavily used in the field of artificial intelligence, especially in robotics [140, 156].

Special cases of the recurrence equations of Proposition 4.1.2 are given, for example, in [140, Section 15.2.1], [156, Section 2.4.3], and [7, Proposition 10.6].

The standard particle filter (also called the bootstrap particle filter or sequential importance resampling particle filter) appeared in [61]. The difficulty of particle filtering in high-dimensional spaces is discussed in [40], [42], and [147], for example.

The conditional particle filter in Section 4.4 is functionally equivalent, but structured in a different way, to the nested particle filter in [31, 32]. However, the discovery of the conditional particle filter came from a different motivation: how to acquire (by filtering) empirical beliefs whose codomains are the space of probability distributions on a structured

space, such as a space of sets, multisets, lists, or function spaces. The presence of these structured spaces suggested deconstructing an empirical belief into two or more levels of 'simpler' empirical beliefs which in turn led to the paired algorithms in Figures 4.17 and 4.21.

For the purposes of tracking states and estimating parameters, the conditional particle filter consists of a pair of mutually recursive filters: one is a particle filter for parameters and the other a particle filter for states that is conditional on the values of the parameter. The parameter filter calls the state filter because it needs to use the current state distribution to approximate the observation likelihood for particular values of the parameters. The state filter calls the parameter filter because it needs to use the parameter particle family approximating the current parameter distribution. Thus the formulation of the conditional particle filter has a clean separation between the parameter and state levels. Furthermore, it can be used for applications beyond the estimation of parameters. The nested particle filter mixes the two levels into a single algorithm, thus arguably missing some conceptual simplicity. Coverage of relevant literature on parameter estimation and the origins of nested particle filters can be found in [32].

In [25], factored conditional filters that simultaneously track states and estimate parameters in high-dimensional state spaces are studied. The conditional nature of the algorithms is used to estimate parameters and the factored nature is used to decompose the state space into low-dimensional subspaces in such a way that filtering on these subspaces gives distributions whose product is a good approximation to the distribution on the entire state space. Ordinary, particle, and variational filters are studied giving twelve algorithms in all. The conditions for successful application of the algorithms are that observations be available at the subspace level and that the transition model can be factored into local transition models that are approximately confined to the subspaces; these conditions are widely satisfied in computer science, engineering, and geophysical filtering applications. Experimental results on tracking epidemics and estimating parameters in large contact networks that show the effectiveness of the approach are given.

An early paper on conditional particle filters in the artificial intelligence literature is [111]. The problem considered is that of simultaneously estimating the pose of a mobile robot and the positions of nearby people in a previously mapped environment. The algorithm proposed to solve the problem is a conditional particle filter that in effect treats the pose of the robot as the parameter space and the position of nearby people as the state space. There are also other meanings of the term conditional particle filter in the literature. For example, in [153], the conditioning is with respect to a state space trajectory.

The idea of factoring appeared in [22, 23], which exploited the structure of a large dynamic Bayesian network (DBN) for efficient approximate inference. The basic idea was to partition the dimension space to obtain tractable approximations of distributions on high-dimensional state spaces. Later, [119] combined factoring with particle filtering to obtain a version of factored particle filters that is similar to the approach in Section 4.5. The main difference between factored particle filters as presented here and the version in [119] is that here the particle family for the entire state space is maintained only in an implicit form; if the state distribution is needed, for example, to compute an integral, particles for the (entire) state are sampled from the implicit form. Later work along these lines includes [34]. The approach here is also similar to that of [40, 41], which use the terminology *multiple* particle filter.

In parallel with the papers above in the artificial intelligence literature, factored particle filters were also studied in the data assimilation literature but using the term *local* particle filter. See the discussion on the origins of localization in [159]. A particularly relevant paper is [134] in which a local particle filter, called a block particle filter, is presented that is similar to the factored nonconditional particle filter in Section 4.5. (A block is the same as a cluster.) This paper also contains a theorem that gives an approximation error bound for the block particle filter that also applies to the factored particle filter here.

Compartmental epidemic models are discussed in [118]. The mathematics of SIS, SEIRS, and other compartmental epidemic models can be found in [74]. Surveys about epidemics on networks can be found in [118, 126, 128].

Exercises

4.1 For an application consisting of a robot in an environment, give concrete examples on an empirical belief, and the corresponding transition and observation models. Would it be feasible for the transition and observation models to be learned during deployment?

4.2 Do Exercise 4.1 again, but this time for applications involving autonomous vehicles, automated trading agents, or personal assistants.

4.3 Explain what changes need to be made to Proposition 4.1.5 to handle the case where there are finitely many empirical beliefs rather than just one.

4.4 For the schema μ and its transition model τ and observation model ξ defined in Example 4.1.4, prove that μ is actually a schema, τ is its transition model, and ξ is its observation model.

4.5 Consider the following schemas, transition models, and observation models:

$$\mu_n^{(1)} : H_n \to \mathcal{P}(X)$$
$$\tau_n^{(1)} : H_{n-1} \times A \times X \to \mathcal{P}(X)$$
$$\xi_n^{(1)} : H_{n-1} \times A \times X \to \mathcal{P}(O)$$

$$\mu_n^{(2)} : H_n \times X \to \mathcal{P}(Y)$$
$$\tau_n^{(2)} : H_{n-1} \times A \times X \times X \times Y \to \mathcal{P}(Y)$$
$$\xi_n^{(2)} : H_{n-1} \times A \times X \times Y \to \mathcal{P}(O)$$

For each $n \in \mathbb{N}$, put

$$\Phi_n \triangleq \lambda(h_{n-1}, a_n, p_n).(h_{n-1}, a_n, p_n, p_n, \mu_{n-1}^{(2)}(h_{n-1}, p_n))$$
$$\Psi_n \triangleq \lambda(h_{n-1}, a_n, p_n, p_{n-1}, \gamma).(h_{n-1}, a_n, p_n, \gamma \odot \lambda y.\tau_n^{(2)}(h_{n-1}, a_n, p_n, p_{n-1}, y))$$
$$\Theta_n \triangleq \lambda(h_{n-1}, a_n, p_n, \gamma).(\gamma \odot \lambda y.\xi_n^{(2)}(h_{n-1}, a_n, p_n, y)).$$

(See Figure 4.33.)

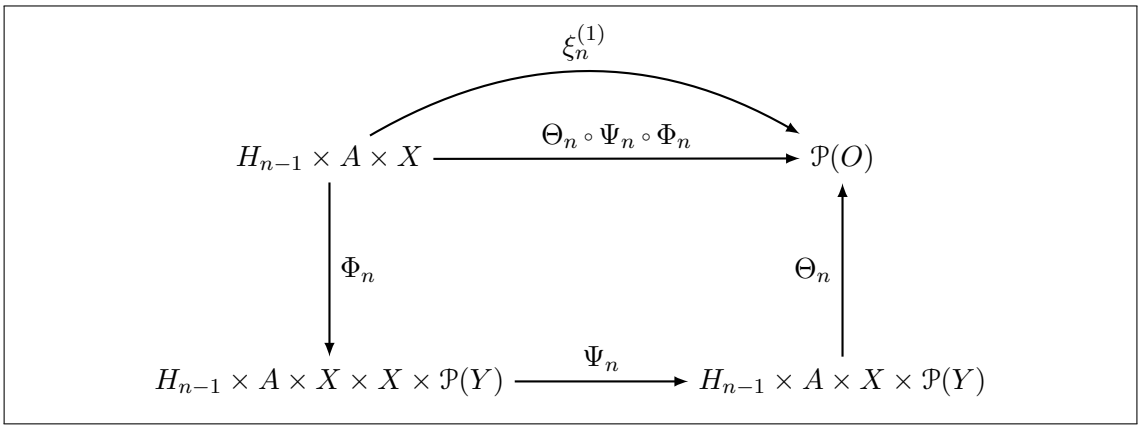

Figure 4.33: Approximation of the observation model for X

1. Prove that the functions

$$\Phi_n : H_{n-1} \times A \times X \to H_{n-1} \times A \times X \times X \times \mathcal{P}(Y)$$
$$\Psi_n : H_{n-1} \times A \times X \times X \times \mathcal{P}(Y) \to H_{n-1} \times A \times X \times \mathcal{P}(Y)$$
$$\Theta_n : H_{n-1} \times A \times X \times \mathcal{P}(Y) \to \mathcal{P}(O)$$

are measurable, for all $n \in \mathbb{N}$.

2. Prove that $\Theta_n \circ \Psi_n \circ \Phi_n : H_{n-1} \times A \times X \to \mathcal{P}(O)$ is a probability kernel, for all $n \in \mathbb{N}$.

3. To what extent could $\Theta_n \circ \Psi_n \circ \Phi_n$ be expected to approximate $\xi_n^{(1)}$?

4.6 Let $(X, \mathcal{A})$ be a measurable space, $f : X \to \mathbb{R}$ a measurable function, and $a \in X$. Prove that $\int_X f \, d\delta_a = f(a)$.

4.7 Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{X})$ a standard Borel space, and $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ a stochastic process. Prove that $\mathbf{x} : \Omega \to X^{\mathbb{N}_0}$ is constant-valued a.s. if and only if $\mathbf{x}_n = \mathbf{x}_{n+1}$ a.s., for all $n \in \mathbb{N}_0$.

Chapter 5

Utilization of Empirical Beliefs

THIS chapter discusses the logicization of beliefs and shows how to reason about beliefs. The first section informally introduces the modal higher-order logic in which the logical form of beliefs is represented. The second section shows how logicization can be achieved. Then follow three sections, one on a variety of examples of computation, one on proof, and one on computation and proof. The sixth section contains examples of reasoning about beliefs, in general. The seventh section is concerned specifically with reasoning about empirical beliefs. Finally, everything is brought together by showing how reasoning about beliefs, especially empirical beliefs, can be exploited by agents to choose appropriate actions to achieve their goals.

5.1 Modal Higher-order Logic

The presentation now widens to consider beliefs more generally, not just empirical beliefs. Recall that a belief is a function that an agent uses to assist in choosing its actions. Some beliefs are built into the agent by the designer, some are acquired by filtering, and some are acquired by other methods. In any case, however an agent acquires its beliefs, to operate it usually needs to reason about its beliefs in order to make good choices of actions. To facilitate this reasoning, it is useful for beliefs to be represented in a logical language, that is, to be logicized. Here, modal higher-order logic is employed. Just the higher-order subset of the logic, without modalities, can be be thought of as a formalization of standard mathematical language. The empirical beliefs of the preceding three chapters and, more generally, beliefs can be represented in modal higher-order logic, that is, logicized. Once beliefs are logicized, sophisticated reasoning about them becomes possible and such reasoning can be exploited to choose actions.

The logicization of beliefs has various subtle aspects, so some explanation of this is needed. Logic has two main aspects, its syntax and its semantics. In the logicization of beliefs, one moves from the semantics to the syntax. A belief is a function — a mathematical object. Mathematical objects live in the semantics of the logic; in particular, they are objects in the intended interpretation of the application. In the syntax, one considers theories that are collection of formulas (or, more generally, biterms). In logicization, a theory is constructed in such a way that the intended interpretation is a model for the theory. This means that, roughly speaking, when interpreted in the intended interpreta-

tion, each formula in the theory is valid. The machinery of reasoning then ensures that any formula that is obtained by reasoning from the theory is a logical consequence of the theory and hence valid in the intended interpretation. Thus logical reasoning produces implicit information that is a consequence of the beliefs of an agent. This information can be used by an agent to assist in choosing its actions.

It is proposed in this book that modal higher-order logic be used to logicize beliefs and reason about them. Note that, importantly, the higher-order subset of the logic (in contrast to, say, first-order logic) is expressive enough to allow the direct representation of probabilistic notions. In other words, one of the advantages of working in a higher-order logic is that it is expressive enough to easily encompass uncertainty without any additional logical machinery.

Higher-order logic admits functions that can take functions as arguments and return functions as results. This feature is heavily exploited by modern declarative programming languages such as Haskell, for which programs consist largely of higher-order functions. For agent applications, two consequences of the use of higher-order logic are noteworthy. One is that there is a class of terms which can be used systematically to represent individuals in applications and this class neatly models (finite) sets and multisets as abstractions (therefore, the functions that process sets and multisets are higher-order). Furthermore, a key idea here that a set is a predicate allows extensional and intensional sets to be processed uniformly by higher-order functions. The other consequence is that the higher-order nature of the logic provides the basis of predicate rewrite systems which are a convenient grammar formalism for expressing hypothesis languages for the belief acquisition facilities of agents.

The use of a logic to represent beliefs opens up another opportunity as well. This is the exploitation of modalities, especially doxastic, epistemic, and temporal ones. Thus the logic employed here is actually *modal* higher-order logic. This means that, for example, it is possible to express beliefs of the form 'agent i believes agent j believes φ' or 'at the last time, agent i believed ψ', where φ and ψ are formulas. Such modalities allow an agent to acquire a highly nuanced model of its environment.

An innovation of the logic is that it allows modalities to be applied to arbitrary terms (not just formulas). So, for example, $\bullet f$, where f is a constant, and $\boldsymbol{B}_i 42$ are legitimate terms. Here, $\bullet$ is the modality 'last time' and $\boldsymbol{B}_i$ is the modality 'agent i believes'. Such terms are called *modal terms*, in contrast with the conventional modal formulas. This extension beyond the standard meaning of modalities will be useful. For example, in machine learning, predicate rewrite systems have been studied before in the case of (non-modal) higher-order logic. That setting is extended here to the modal case by allowing modalities to be applied to predicates. As another example, for belief acquisition, it is useful to allow modalities to be applied to constants.

Another feature of the logic is the polymorphic type system. It is not necessary to say much about the need for a type system. Apart from other considerations, good software engineering practice demands a type system to accurately model applications and avoid many typical programming errors. (Parametric) polymorphism, in particular, is also needed for many common functions that operate in a similar way on a variety of data types. A programming language such as Haskell provides a good illustration of the value of having a sophisticated type system.

As well as representing beliefs, it is necessary to reason about them. The reasoning system employed here combines a computation component and a proof component. The

theorem prover is a fairly conventional tableau theorem prover for modal higher-order logic. The computation component significantly extends existing declarative programming languages by adding facilities for computing with modalities. The proof component and the computational component are tightly integrated, in the sense that either can call the other. Furthermore, this synergy between the two is shown to make possible all kinds of interesting reasoning tasks. It turns out that, for agent applications, the most common reasoning task is a computational one, that of evaluating a function call. In this case, the theorem-prover plays a subsidiary role, usually that of performing some rather straightforward modal theorem-proving tasks. However, in other applications it can just as easily be the other way around with the computational system performing subsidiary equational reasoning tasks for the theorem prover. The theoretical foundations for the reasoning system that are developed in the book support both cases. The presentation below of the reasoning system considers first the case of (pure) computation, where no proof is involved, then (pure) proof, where no computation is involved, and finally the two are put together.

In many applications, it is necessary to deal with uncertainty. One of the great advantages of working in a higher-order logic is that it is expressive enough to easily encompass uncertainty without any additional machinery. The key idea is to represent uncertainty by probabilities densities. Densities can be conveniently included in theories and manipulated by higher-order functions. In this way, it is straightforward to represent empirical beliefs directly in a theory and reason about them.

The reasoning system described here is embodied in the Bach programming language [101]. Being based on modal higher-order logic, Bach is highly expressive for knowledge representation. Since higher-order logic is the formalization of standard mathematical language, it follows that every application that can be modelled in standard mathematical language can be modelled directly in Bach. Modalities add extra expressive power, while uncertainty in the form of probability densities can be modelled directly in the language.

A brief summary of the syntax of the logic is now given. Types and terms are defined, and an introduction is given to the modalities that will be used. Full details of the logic can be found in Appendices B.1, B.2, and B.3.

An *alphabet* consists of three sets: a set $\mathfrak{T}$ of type constructors; a set $\mathfrak{C}$ of constants; and a set $\mathfrak{V}$ of variables.

Each type constructor in $\mathfrak{T}$ has an arity. The set $\mathfrak{T}$ always includes the type constructor *Bool* of arity 0. *Bool* is the type of the booleans. Each constant in $\mathfrak{C}$ has a signature. The set $\mathfrak{V}$ is denumerable. Variables are typically denoted by $x, y, z, \ldots$.

Types are built up from the set of type constructors using the symbols $\rightarrow$ and $\times$.

A *type* is defined inductively as follows.

1. If T is a type constructor of arity k and $\alpha_1, \ldots, \alpha_k$ are types, then $T\ \alpha_1 \ldots \alpha_k$ is a type. (Thus a type constructor of arity 0 is a type.)

2. If α and β are types, then $\alpha \rightarrow \beta$ is a type.

3. If $\alpha_1, \ldots, \alpha_n$ are types, then $\alpha_1 \times \cdots \times \alpha_n$ is a type.

Example 5.1.1. Following are some common types other than *Bool* that will be needed. The type of the integers is denoted by *Int*, and the type of the reals by *Real*. Also (*List* σ) is

the type of lists whose items have type σ. Here, *Int*, *Real* and *List* are all type constructors. The first two have arity 0 and the last has arity 1. A function that maps elements of type α to elements of type β has type $\alpha \to \beta$. Since sets are identified with predicates in the logic, sets whose elements have type σ have type $\sigma \to \mathit{Bool}$. Sometimes $\{\sigma\}$ is written as a synonym for $\sigma \to \mathit{Bool}$ when it is helpful to make a distinction between sets and predicates. A particular class of functions of interest is that of probability densities. The synonym $\mathit{Density}\ \tau \triangleq \tau \to \mathit{Real}$ is introduced, but with the understanding that the meaning of a term of type $\mathit{Density}\ \tau$ is a probability density over elements of type τ rather than an arbitrary real-valued function over elements of type τ.

The set $\mathfrak{C}$ always includes the following constants.

1. $\top$ and $\bot$, having signature *Bool*.
2. $=_\alpha$, having signature $\alpha \to \alpha \to \mathit{Bool}$, for each type α.
3. $\neg$, having signature $\mathit{Bool} \to \mathit{Bool}$.
4. $\wedge$, $\vee$, and $\longrightarrow$ having signature $\mathit{Bool} \to \mathit{Bool} \to \mathit{Bool}$.
5. Σ_α and Π_α, having signature $(\alpha \to \mathit{Bool}) \to \mathit{Bool}$, for each type α.

The intended meaning of $\top$ is true, and that of $\bot$ is false. The intended meaning of $=_\alpha$ is identity, and the intended meanings of the connectives $\neg$, $\wedge$, $\vee$, and $\longrightarrow$ are as usual. The intended meanings of Σ_α and Π_α are as follows: Σ_α maps a predicate to $\top$ iff the predicate maps at least one element to $\top$; Π_α maps a predicate to $\top$ iff the predicate maps all elements to $\top$.

Other useful constants that will commonly appear in applications include the integers, the real numbers, and data constructors like $\sharp_\sigma : \sigma \to \mathit{List}\ \sigma \to \mathit{List}\ \sigma$ and $[]_\sigma : \mathit{List}\ \sigma$ for constructing lists with elements of type σ. The notation $C : \sigma$ is used to indicate that the constant C has signature σ.

Necessity modalities $\Box_i$, for $i = 1, \ldots, m$, are assumed.

A *term*, together with its type, is defined inductively as follows.

1. A variable in $\mathfrak{V}$ of type α is a term of type α.
2. A constant in $\mathfrak{C}$ having signature α is a term of type α.
3. (Abstraction) If t is a term of type β and x a variable of type α, then $\lambda x.t$ is a term of type $\alpha \to \beta$.
4. (Application) If s is a term of type $\alpha \to \beta$ and t a term of type α, then $(s\ t)$ is a term of type β.
5. (Tuple) If $t_1, \ldots, t_n$ are terms of type $\alpha_1, \ldots, \alpha_n$, respectively, then $(t_1, \ldots, t_n)$ is a term of type $\alpha_1 \times \cdots \times \alpha_n$.
6. (Modal Term) If t is a term of type α and $i \in \{1, \ldots, m\}$, then $\Box_i t$ is a term of type α.

Example 5.1.2. Constants like $\top : \mathit{Bool}$, $42 : \mathit{Int}$, $3.11 : \mathit{Real}$, and $+ : \mathit{Int} \to \mathit{Int} \to \mathit{Int}$ are terms. Variables like x, y, and z are terms. An example of a term that can be formed using abstraction is $\lambda x.((+\ x)\ x)$ of type $\mathit{Int} \to \mathit{Int}$, whose intended meaning is a function that takes a number x and returns $x + x$. To apply that function to the constant 42, for example, application is used to form the term $(\lambda x.((+\ x)\ x)\ 42)$, which has type Int.

Example 5.1.3. The term $(\sharp_{\mathit{Int}}\ 2\ (\sharp_{\mathit{Int}}\ 3\ []_{\mathit{Int}}))$ of type $(\mathit{List}\ \mathit{Int})$ represents a list with the integers 2 and 3 in it, obtained via a series of applications from the constants $\sharp_{\mathit{Int}}, []_{\mathit{Int}}, 2$, and 3, each of which is a term. For convenience, $[2, 3]$ is sometimes written to represent the same list.

Example 5.1.4. Sets are identified with predicates in the logic. Thus, the term

$$\lambda x.((\vee\ ((=_{\mathit{Int}}\ x)\ 2))\ ((=_{\mathit{Int}}\ x)\ 3)) \tag{5.1.1}$$

of type $\mathit{Int} \to \mathit{Bool}$ can be used to represent the set containing the integers 2 and 3. Infix notation is often used for common constants like equality and the connectives. The convention is also adopted that applications are left-associative; thus $(f\ x\ y)$ means $((f\ x)\ y)$. These conventions allow us to write $\lambda x.((x =_{\mathit{Int}} 2) \vee (x =_{\mathit{Int}} 3))$ instead of (5.1.1) above. For convenience, $\{2, 3\}$ is sometimes also written to represent the same set. Since sets are predicates, set membership test is obtained using function application. Let s denote (5.1.1) above. To check whether a number y is in the set, one just writes $(s\ y)$.

Terms of the form $(\Sigma_\alpha\ \lambda x.t)$ are written as $\exists_\alpha x.t$ and terms of the form $(\Pi_\alpha\ \lambda x.t)$ are written as $\forall_\alpha x.t$ (in accord with the intended meaning of Σ_α and Π_α). A formula is a term of type Bool. The universal closure of a formula φ is denoted by $\forall(\varphi)$.

There is a *default term* for each type. For example, the default term of type Bool is $\bot$, that of type Int is 0, that of type $\mathit{List}\ \alpha$ for any α is $[]_\alpha$, and that of type $\{\alpha\}$ for any α is $\{\}$ (that is, $\lambda x.\bot$).

The polymorphic version of the logic extends what is given above by also having available parameters which are type variables (denoted by $a, b, c, \ldots$). The definition of a type as above is then extended to polymorphic types that may contain parameters and the definition of a term as above is extended to terms that may have polymorphic types. The polymorphic version of the logic is employed in the remainder of this chapter. In this case, the α is dropped in constants like $\exists_\alpha$, $\forall_\alpha$, $=_\alpha$, $[]_\alpha$ and $\sharp_\alpha$, since the types associated with these are now inferred from the context.

Example 5.1.5. A common polymorphic constant needed is $\mathit{if\_then\_else} : \mathit{Bool} \times a \times a \to a$. Using it, one can give the following equivalent way of writing (5.1.1) above:

$$\lambda x.(\mathit{if\_then\_else}\ ((x = 2), \top, (\mathit{if\_then\_else}\ ((x = 3), \top, \bot)))).$$

Writing $\mathit{if}\ x\ \mathit{then}\ y\ \mathit{else}\ z$ as syntactic sugar for $(\mathit{if\_then\_else}\ (x, y, z))$, the above can be written in the following more readable form:

$$\lambda x.\mathit{if}\ x = 2\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = 3\ \mathit{then}\ \top\ \mathit{else}\ \bot.$$

Discrete probability densities can also be written down easily as terms using $\mathit{if\_then\_else}$. For instance, the term

$$\lambda x.\mathit{if}\ x = \top\ \mathit{then}\ 0.3\ \mathit{else}\ \mathit{if}\ x = \bot\ \mathit{then}\ 0.7\ \mathit{else}\ 0$$

of type $\mathit{Density}\ \mathit{Bool}$ denotes a probability density over the booleans.

Modalities can have a variety of meanings, depending on the application. Some of these are now indicated. The setting here is more general than usual in that it allows modalities to be applied to terms and dual modalities to be applied to biterms (Definition B.1.10). For example, consider an application with three agents. One meaning for the necessity modality is knowledge. So, $\Box_i\varphi$, for $i = 1$, 2, and 3, is used to to denote 'agent i knows φ'. In this case, the modalities $\Box_1$, $\Box_2$, and $\Box_3$ can be more meaningfully written as $\boldsymbol{K}_1$, $\boldsymbol{K}_2$, and $\boldsymbol{K}_3$. A modality weaker than knowledge is that of belief. One can use $\Box_i\varphi$, for $i = 4$, 5, and 6, to denote 'agent $(i-3)$ believes φ'. In this case, $\Box_4$, $\Box_5$, and $\Box_6$ can be written as $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, and $\boldsymbol{B}_3$. Modalities can also have a variety of temporal readings. One can introduce $\Box_7$ for 'next' (written as $\bigcirc$), $\Box_8$ for 'always in the future' (written simply as $\Box$), $\Box_9$ for 'last' (written as $\bullet$), and $\Box_{10}$ for 'always in the past' (written as $\blacksquare$). Taking the dual of $\Box$ and $\blacksquare$, $\Diamond$ ('sometime in the future') and $\blacklozenge$ ('sometime in the past') are obtained. Modal terms such as $\boldsymbol{B}_i 42$ and $\bullet f$, where f is a constant having signature of the form $\sigma \to \tau$, are admitted. The need for modal terms arises naturally in applications, as shall be seen below.

The logic can be given a rather conventional semantics in the usual Kripke style for modal logics, with higher-order interpretations at each world. However, since the concept of a modal term is new in modal logic, some intuition for the semantics of modal terms is given. If t is a formula, then the meaning of $\Box_i t$ in a world is T if the meaning of t in all accessible worlds is T, its meaning is F if the meaning of t in all accessible worlds is F, and, in the other cases, the meaning of $\Box_i t$ is conventionally defined to be F. This suggests an obvious extension to terms t that have rank 0 (that is, do not have type of the form $\alpha \to \beta$): if t has the same meaning in all accessible worlds, then the meaning of $\Box_i t$ should be this common meaning; otherwise, the meaning of $\Box_i t$ should be some default value. This definition then becomes the base case of an inductive definition on the rank of the type of t of the semantics of a modal term $\Box_i t$.

Each application has a distinguished pointed interpretation (I, w) known as the *intended pointed interpretation*, where I is an interpretation and w is a world in I. This means that, in the application, w is the actual world and I provides the worlds accessible to w by the various accessibility relations.

In modal logics, constants generally have different meanings in different worlds. Certain constants can be declared to be rigid; they then have the same meaning in all worlds (in the semantics). Except in the most sophisticated applications, it is entirely natural for some constants to be rigid. For instance, one can declare all data constructors (e.g. $\top, \bot, 1, 2, 3, \ldots, \sharp, []$) to be rigid. Also, all constants in the Haskell Prelude, which is a library of basic function definitions, can be declared to be rigid. A term is *rigid* if every constant in it is rigid.

A theory, which is a set of formulas, can consist of two kinds of assumptions, global and local. The essential difference is that global assumptions are true in each world in the intended pointed interpretation, while local assumptions only have to be true in the actual world in the intended pointed interpretation. Each kind of assumption has a certain role to play in computations. A theory is denoted by a pair $(\mathcal{G}, \mathcal{L})$, where $\mathcal{G}$ is the set of global assumptions and $\mathcal{L}$ is the set of local assumptions.

Typically, for agent i, local assumptions in its belief theory have the form $\boldsymbol{B}_i\varphi$, with the intuitive meaning 'agent i believes φ'. Other typical local assumptions have the form $\boldsymbol{B}_i\boldsymbol{B}_j\varphi$, meaning 'agent i believes that agent j believes φ'. Global assumptions in a belief

theory typically have the form φ, with no modalities at the front since the fact that they are global implicitly implies any sequence of (necessity) modalities effectively appears at the front. If there is a temporal component to belief formulas, this is often manifested by temporal modalities at the front of belief formulas. Then, for example, there could be a belief formula of the form $\bullet^2 \boldsymbol{B}_i \boldsymbol{B}_j \varphi$, whose intuitive meaning is 'at the second last time, agent i believed that agent j believed φ'. (Here, $\bullet^2$ is a shorthand for $\bullet\bullet$.) Thus belief formulas commonly have the form $\Box_{j_1} \cdots \Box_{j_n} \varphi$, where $n \geq 0$.

In multi-agent applications, one meaning for $\Box_i t$ is that 'agent i knows t', where t is a term. In this case, the modality $\Box_i$ is written as $\boldsymbol{K}_i$. The logic $\mathbf{S5}_m$ is commonly used to capture the intended meaning of knowledge. The axiom schemes for this logic are as follows (where $i = 1, \ldots, m$).

$$\boldsymbol{K_i}(\boldsymbol{\varphi} \longrightarrow \boldsymbol{\psi}) \longrightarrow (\boldsymbol{K_i\varphi} \longrightarrow \boldsymbol{K_i\psi}) \tag{Distr$_i$}$$

$$\boldsymbol{K}_i\boldsymbol{\varphi} \longrightarrow \boldsymbol{\varphi} \tag{T_i}$$

$$\boldsymbol{\varphi} \longrightarrow \boldsymbol{K}_i\neg\boldsymbol{K}_i\neg\boldsymbol{\varphi} \tag{B_i}$$

$$\boldsymbol{K}_i\boldsymbol{\varphi} \longrightarrow \boldsymbol{K}_i\boldsymbol{K}_i\boldsymbol{\varphi} \tag{4_i}$$

where $\boldsymbol{\varphi}$ is a syntactical variable ranging over biterms. The first axiom (scheme) is the distribution axiom and is valid in every interpretation. (See Proposition B.2.21). The accessibility relation corresponding to axiom (T_i) is reflexivity; for axiom (B_i), it is symmetry; and for axiom (4_i), it is transitivity. (See Proposition B.2.41.)

Given the form of axiom (B_i), it is helpful to introduce some new notation and terminology. Let $\boldsymbol{E}_i \triangleq \neg\boldsymbol{K}_i\neg$. Thus $\boldsymbol{E}_i\varphi \triangleq \neg\boldsymbol{K}_i\neg\varphi$, for all biterms φ. The meaning of $\boldsymbol{E}_i\varphi$ is 'agent i entertains φ', in the sense that the agent entertains the possibility that φ is true. (More literally, the meaning is that 'it is not true that agent i knows that φ is false'.) Thus axiom (B_i) can be written in the form

$$\boldsymbol{\varphi} \longrightarrow \boldsymbol{K}_i\boldsymbol{E}_i\boldsymbol{\varphi}.$$

Informally, the distribution axiom for knowledge (Distr$_i$) states that if agent i knows that φ implies ψ and agent i knows φ, then agent i knows ψ. Axiom (T_i) states that if agent i knows φ, then φ is true. Axiom (B_i) states that if φ is true, then agent i knows that agent i entertains φ. Axiom (4_i) states that if agent i knows φ, then agent i knows that agent i knows φ.

Now the discussion turns to beliefs. In this case, $\Box_i t$ means that 'agent i believes t' and the modality $\Box_i$ is written as $\boldsymbol{B}_i$. Consider the logicization $\boldsymbol{B}_i\varphi$ of a belief of agent i. Then $\boldsymbol{B}_i\varphi$ is valid in the intended pointed interpretation; however, in contrast to knowledge, φ may not be valid in the intended pointed interpretation. Informally, a belief of an agent, especially an empirical belief, may not be 'true', although it can be hoped that it provides a good approximations to the truth. Put another way, it is true that the agent i believes φ, but φ itself may not be true.

In more detail, suppose that φ and ψ are (closed) formulas and (I, w) the intended pointed interpretation. Then it may happen that $\mathcal{V}(\boldsymbol{B}_i\varphi, I, w) = \mathsf{T}$ ("agent i believes φ") and $\mathcal{V}(\boldsymbol{B}_i\boldsymbol{B}_j\psi, I, w) = \mathsf{T}$ ("agent i believes agent j believes φ"), but $\mathcal{V}(\varphi, I, w) = \mathsf{F}$ ("φ is false") and $\mathcal{V}(\psi, I, w) = \mathsf{F}$ ("ψ is false"). Axiom (T_i) is thus problematical for beliefs. Hence axioms for the notion of belief are needed that are weaker than the axioms for knowledge.

The logic $\mathbf{KD45}_m$ is commonly used to capture the intended meaning of belief. The axiom schemes for this logic are as follows (where $i = 1, \ldots, m$).

$$\boldsymbol{B_i}(\boldsymbol{\varphi} \longrightarrow \boldsymbol{\psi}) \longrightarrow (\boldsymbol{B_i\varphi} \longrightarrow \boldsymbol{B_i\psi}) \tag{Distr$_i$}$$
$$\boldsymbol{B}_i\boldsymbol{\varphi} \longrightarrow \neg\boldsymbol{B}_i\neg\boldsymbol{\varphi} \tag{D_i}$$
$$\boldsymbol{B}_i\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_i\boldsymbol{B}_i\boldsymbol{\varphi} \tag{4_i}$$
$$\neg\boldsymbol{B}_i\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_i\neg\boldsymbol{B}_i\boldsymbol{\varphi} \tag{5_i}$$

where $\boldsymbol{\varphi}$ is a syntactical variable ranging over biterms. The accessibility relation corresponding to axiom (D_i) is seriality (that is, for all $w \in W$, there exists $w' \in W$ such that $w\, R_i\, w'$); for axiom (4_i), it is transitivity; and for axiom (5_i), it is Euclideaness (that is, for all $w, w', w'' \in W$, $w\, R_i\, w'$ and $w\, R_i\, w''$ implies $w'\, R_i\, w''$). (See Proposition B.2.41.)

Given the form of axioms (D_i) and (5_i), it is helpful to introduce some new notation and terminology. Let $\boldsymbol{C}_i \triangleq \neg\boldsymbol{B}_i\neg$. Thus $\boldsymbol{C}_i\varphi \triangleq \neg\boldsymbol{B}_i\neg\varphi$, for all biterms φ. The meaning of $\boldsymbol{C}_i\varphi$ is 'agent i contemplates φ', in the sense that the agent contemplates the possibility that φ is true. Thus axiom (D_i) can be written in the form

$$\boldsymbol{B}_i\boldsymbol{\varphi} \longrightarrow \boldsymbol{C}_i\boldsymbol{\varphi}$$

axiom (5_i) can be written in the form

$$\boldsymbol{C}_i\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_i\boldsymbol{C}_i\boldsymbol{\varphi}.$$

Informally, the distribution axiom for belief (Distr$_i$) states that if agent i believes that φ implies ψ and agent i believes φ, then agent i believes ψ. Axiom (D_i) states that if agent i believes φ, then agent contemplates φ. Axiom (4_i) states that if agent i believes φ, then agent i believes that agent i believes φ. Axiom (5_i) states that if agent i contemplates φ, then agent i believes that agent i contemplates φ. The crucial difference between the logic $\mathbf{S5}_m$ of knowledge and the logic $\mathbf{KD45}_m$ of belief is that the latter logic avoids use of the axiom (T_i).

A term of the form $\boldsymbol{B}_i t$, where t is not a formula, while perhaps unfamiliar, is intuitive and useful. Intuitively, $\boldsymbol{B}_i t$ means 'agent i believes t'; formally, its semantics is given by the $\mathcal{M}$ function in Definition B.2.10. A typical example of this, other than when t is a formula, is when t is a constant, say, f having signature of the form $\sigma \to \tau$. Then $\boldsymbol{B}_i f$ means 'agent i believes function f'. This is normally captured by a definition for f that is the one believed by agent i (in contrast to other definitions for f that may be believed by other agents).

There is considerable attention in the literature on logics such as $\mathbf{S5}_m$, $\mathbf{KD45}_m$, and related ones concerning the subtle differences in meaning that the various axioms might bestow on modalities for knowledge and belief. When trying to build agent applications these subtleties can be overshadowed by the issues of dealing with the various function definitions for different agents at different times and the choice of interaction axioms that relate the various modalities. In any case, the consequences that might follow from using the axioms for the logics $\mathbf{S5}_m$ and $\mathbf{KD45}_m$ are not explored in detail in this book.

The modalities also have a variety of temporal readings. Here, one version of the case of discrete time is remarked upon. For this case, the intended interpretation includes the

natural numbers $\mathbb{N}_0$ (or, perhaps, the integers $\mathbb{Z}$). Each $n \in \mathbb{N}_0$ is interpreted as a time point with $n+1$ the next time point after n.

The usual modalities $\bigcirc$ ('next'), $\Box$ ('always in the future'), $\Diamond$ ('sometime in the future'), and $\boldsymbol{U}$ ('until') are adopted. The modalities $\bigcirc$ and $\Box$ are assumed to be among the $\Box_1, \ldots, \Box_m$. In particular, $\bigcirc$ is interpreted by the relation $\{(n, n+1) \mid n \in \mathbb{N}_0\}$. The modalities $\bigcirc$ and $\Box$ can be applied to arbitrary terms. The modality $\Diamond$ is dual to $\Box$ and thus can only be applied to biterms.

The modality $\boldsymbol{U}$ is a binary modality and thus, strictly speaking, cannot even be written down in the logic since it only admits unary modalities. However, as will now be shown, one can introduce $\boldsymbol{U}$ as an abbreviation for a higher-order term involving $\bigcirc$ that can be applied to biterms. By this means, the higher-order nature of the logic be used to effectively admit a binary modality into the logic. For this, $\bigcirc^n$ and $\bigcirc^{\leq n}$, where $n \in \mathbb{N}_0$, need to be defined. Let φ be a biterm. Then $\bigcirc^0\varphi = \bigcirc^{\leq 0}\varphi = \varphi$ and, for $n \geq 0$,

$$\bigcirc^{n+1}\varphi = \bigcirc\bigcirc^n\varphi$$
$$\bigcirc^{\leq n+1}\varphi = \bigwedge_{i=0}^{n+1} \bigcirc^i\varphi.$$

These definitions can be expressed in the logic by the following axiom schemes.

$$\mathit{sometime} : \alpha \times \mathit{Nat} \to \alpha$$
$$(\mathit{sometime}\ (\boldsymbol{\varphi}, n)) = \mathit{if}\ n = 0\ \mathit{then}\ \boldsymbol{\varphi}\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\boldsymbol{\varphi}, n-1))$$

$$\mathit{upto\_sometime} : \alpha \times \mathit{Nat} \to \alpha$$
$$(\mathit{upto\_sometime}\ (\boldsymbol{\varphi}, n)) =$$
$$\mathit{if}\ n = 0\ \mathit{then}\ \boldsymbol{\varphi}\ \mathit{else}\ ((\mathit{sometime}\ (\boldsymbol{\varphi}, n)) \wedge (\mathit{upto\_sometime}\ (\boldsymbol{\varphi}, n-1))),$$

where α is a biterm type and $\boldsymbol{\varphi}$ is a syntactical variable standing for an arbitrary biterm of type α. (Syntactical variables are needed (at least in the definition of *sometime*) because the modality restricts too much the substitutions that can be made in the first argument of *sometime*.) Then, for $n \in \mathbb{N}_0$ and φ a biterm, the procedure of Section B.3.1 can be used to compute $(\mathit{sometime}\ (\varphi, n))$ and $(\mathit{upto\_sometime}\ (\varphi, n))$. The answers returned are

$$(\mathit{sometime}\ (\varphi, n)) = \bigcirc^n\varphi$$

and

$$(\mathit{upto\_sometime}\ (\varphi, n)) = \bigcirc^{\leq n}\varphi.$$

Now $\varphi\,\mathbf{U}\,\psi$ can be defined with the meaning 'φ holds until ψ holds'. Define $\varphi\,\boldsymbol{U}\,\psi$ to be an abbreviation for

$$\psi \vee \exists n.((\mathit{upto\_sometime}\ (\varphi, n)) \wedge (\mathit{sometime}\ (\psi, n+1))).$$

This clearly captures the intended meaning of $\boldsymbol{U}$.

Next are given a useful collection of equations about the modalities $\bigcirc$, $\Box$, $\Diamond$, and $\boldsymbol{U}$ that are valid at every world in every interpretation and, therefore, can be used as global assumptions in any theory.

$$\begin{aligned}
&\boldsymbol{\varphi}\,\boldsymbol{U}\,\boldsymbol{\psi} = \boldsymbol{\psi} \vee (\boldsymbol{\varphi} \wedge \bigcirc(\boldsymbol{\varphi}\,\boldsymbol{U}\,\boldsymbol{\psi}))\\
&\Diamond\boldsymbol{\varphi} = \boldsymbol{\varphi} \vee \bigcirc\Diamond\boldsymbol{\varphi}\\
&\Box\boldsymbol{\varphi} = \boldsymbol{\varphi} \wedge \bigcirc\Box\boldsymbol{\varphi}\\
&\bigcirc(\boldsymbol{\varphi} \wedge \boldsymbol{\psi}) = \bigcirc\boldsymbol{\varphi} \wedge \bigcirc\boldsymbol{\psi}\\
&\bigcirc(\boldsymbol{\varphi} \vee \boldsymbol{\psi}) = \bigcirc\boldsymbol{\varphi} \vee \bigcirc\boldsymbol{\psi}\\
&\bigcirc(\boldsymbol{\varphi} \longrightarrow \boldsymbol{\psi}) = \bigcirc\boldsymbol{\varphi} \longrightarrow \bigcirc\boldsymbol{\psi}\\
&\bigcirc\neg\boldsymbol{\varphi} = \neg\bigcirc\boldsymbol{\varphi},
\end{aligned}$$

where $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are syntactical variables ranging over biterms of the same type.

Companions to the above (future) temporal modalities are the past temporal modalities $\bullet$ ('at the last time'), $\blacksquare$ ('always in the past'), $\blacklozenge$ ('sometime in the past'), and $\boldsymbol{S}$ ('since'). The modality $\bullet$ is interpreted by the relation $\{(n+1, n) \mid n \in \mathbb{N}_0\} \cup \{(0,0)\}$. Also $\boldsymbol{S}$ can be defined in terms of $\bullet$ in an analogous way as for the corresponding definitions for the future temporal modalities. There is also a set of equations for the past temporal modalities that are useful for reasoning.

$$\begin{aligned}
&\boldsymbol{\varphi}\,\boldsymbol{S}\,\boldsymbol{\psi} = \boldsymbol{\psi} \vee (\boldsymbol{\varphi} \wedge \bullet(\boldsymbol{\varphi}\,\boldsymbol{S}\,\boldsymbol{\psi}))\\
&\blacklozenge\boldsymbol{\varphi} = \boldsymbol{\varphi} \vee \bullet\blacklozenge\boldsymbol{\varphi}\\
&\blacksquare\boldsymbol{\varphi} = \boldsymbol{\varphi} \wedge \bullet\blacksquare\boldsymbol{\varphi}\\
&\bullet(\boldsymbol{\varphi} \wedge \boldsymbol{\psi}) = \bullet\boldsymbol{\varphi} \wedge \bullet\boldsymbol{\psi}\\
&\bullet(\boldsymbol{\varphi} \vee \boldsymbol{\psi}) = \bullet\boldsymbol{\varphi} \vee \bullet\boldsymbol{\psi}\\
&\bullet(\boldsymbol{\varphi} \longrightarrow \boldsymbol{\psi}) = \bullet\boldsymbol{\varphi} \longrightarrow \bullet\boldsymbol{\psi}\\
&\bullet\neg\boldsymbol{\varphi} = \neg\bullet\boldsymbol{\varphi},
\end{aligned}$$

where $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are syntactical variables ranging over biterms of the same type.

The next global assumption is given by Proposition B.2.15 which shows that the following scheme can be used as a global assumption.

$$\Box_i\boldsymbol{t} = \boldsymbol{t}, \tag{5.1.2}$$

where $\boldsymbol{t}$ is a syntactical variable ranging over *rigid* terms and $i \in \{1, \ldots, m\}$. Instances of this scheme that could be used as global assumptions include the following.

$$\begin{aligned}
&\boldsymbol{B}_i 42 = 42\\
&\bullet 42 = 42\\
&\boldsymbol{B}_i[] = []\\
&\boldsymbol{B}_i\top = \top\\
&\boldsymbol{B}_i\bot = \bot\\
&\bullet\top = \top\\
&\bullet\bot = \bot.
\end{aligned}$$

Proposition B.2.18 shows that the following scheme dual to (5.1.2) can be used as a global assumption.

$$\Diamond_i \boldsymbol{\varphi} = \boldsymbol{\varphi}, \tag{5.1.3}$$

where $\boldsymbol{\varphi}$ is a syntactical variable ranging over *rigid* biterms and $i \in \{1, \ldots, m\}$.

Proposition B.2.16 shows that the following scheme can be used as a global assumption.

$$(\Box_i \mathbf{s}\ \mathbf{t}) = \Box_i(\mathbf{s}\ \mathbf{t}), \tag{5.1.4}$$

where $\mathbf{s}$ is a syntactical variable ranging over terms having type of the form $\alpha \to \beta$, $\mathbf{t}$ is a syntactical variable ranging over *rigid* terms of type α, and $i \in \{1, \ldots, m\}$.

Specialized to some of the epistemic and temporal modalities discussed so far, this means that

$$(\boldsymbol{B}_i \mathbf{s}\ \mathbf{t}) = \boldsymbol{B}_i(\mathbf{s}\ \mathbf{t})$$
$$(\bigcirc \mathbf{s}\ \mathbf{t}) = \bigcirc(\mathbf{s}\ \mathbf{t})$$
$$(\Box \mathbf{s}\ \mathbf{t}) = \Box(\mathbf{s}\ \mathbf{t})$$
$$(\bullet \mathbf{s}\ \mathbf{t}) = \bullet(\mathbf{s}\ \mathbf{t})$$
$$(\blacksquare \mathbf{s}\ \mathbf{t}) = \blacksquare(\mathbf{s}\ \mathbf{t})$$

are global assumptions (under the rigidity assumption on $\mathbf{t}$).

Global assumption (5.1.4) is also useful when the rank of the type of s is > 1. Suppose that s is a term whose type has rank 2 and t_1 and t_2 are rigid terms such that $((s\ t_1)\ t_2)$ is a term. Then $((\Box_i s\ t_1)\ t_2)$ can be rewritten first to $(\Box_i(s\ t_1)\ t_2)$ and then to $\Box_i((s\ t_1)\ t_2)$.

Proposition B.2.19 shows that the following scheme dual to (5.1.4) can be used as a global assumption.

$$(\Diamond_i \boldsymbol{\varphi}\ \mathbf{t}) = \Diamond_i(\boldsymbol{\varphi}\ \mathbf{t}), \tag{5.1.5}$$

where $\boldsymbol{\varphi}$ is a syntactical variable ranging over biterms having type of the form $\alpha \to \beta$, $\mathbf{t}$ is a syntactical variable ranging over *rigid* terms of type α, and $i \in \{1, \ldots, m\}$.

Proposition B.2.17 shows that the following scheme can be used as a global assumption.

$$\Box_i \lambda x.\mathbf{t} = \lambda x.\Box_i \mathbf{t}, \tag{5.1.6}$$

where $\mathbf{t}$ is a syntactical variable ranging over terms and $i \in \{1, \ldots, m\}$.

Proposition B.2.20 shows that the following scheme dual to (5.1.6) can be used as a global assumption.

$$\Diamond_i \lambda x.\boldsymbol{\varphi} = \lambda x.\Diamond_i \boldsymbol{\varphi}, \tag{5.1.7}$$

where $\boldsymbol{\varphi}$ is a syntactical variable ranging over biterms and $i \in \{1, \ldots, m\}$.

In addition, there are interactions between the temporal and epistemic modalities for which implicational schemes such as the following may be appropriate (depending on the application). The following interaction axioms model agents with perfect recall [47, p.286].

$$\boldsymbol{B}_i \bigcirc \boldsymbol{\varphi} \longrightarrow \bigcirc \boldsymbol{B}_i \boldsymbol{\varphi}$$
$$\boldsymbol{K}_i \bigcirc \boldsymbol{\varphi} \longrightarrow \bigcirc \boldsymbol{K}_i \boldsymbol{\varphi},$$

where $\boldsymbol{\varphi}$ is a syntactical variable ranging over biterms.

Similarly, there are implicational schemes available for use that connect the past temporal and epistemic modalities. Consider following the interaction axioms that model another form of perfect recall.

$$\bullet \boldsymbol{B}_i \boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_i \bullet \boldsymbol{\varphi}$$
$$\bullet \boldsymbol{K}_i \boldsymbol{\varphi} \longrightarrow \boldsymbol{K}_i \bullet \boldsymbol{\varphi},$$

where $\boldsymbol{\varphi}$ is a syntactical variable ranging over biterms. In Section 5.3, an illustration of a computation using the past temporal modalities will be given. Generally, the past temporal modalities are useful in applications for which an agent uses the past (and present) to select a suitable action.

5.2 Logicization

What the methods of Chapters 3 and 4 produce is definitions of functions that are empirical beliefs. These are concrete functions that form part of the intended interpretation. However, so far, this intended interpretation is not sufficiently formalized to be able to carry out the next stage which is that of constructing a suitable theory about the empirical beliefs for which the intended interpretation is a model. What is missing is the statement of the other components of an interpretation in the logic. (See Definition B.2.5.) So the set of worlds, relations, domains, and so on, need to be specified, at least to the extent that the transition from the intended interpretation to the theory can be carried out for the set of empirical beliefs. And, of course, this has to be done at each time step, since empirical beliefs in general change as new observations are made.

Leaving aside the modalities for the moment, the translation of a an empirical belief into a formula in the logic is comparatively straightforward. But getting the correct modalities in the correct places in this formula is more complicated.

Having introduced the logic, the logicization of beliefs is explained. For concreteness, consider a belief $\mathbf{f} : X \to Y$ that is defined by an expression of the form

$$\forall \mathbf{x}.(\mathbf{f}(\mathbf{x}) = \mathbf{t}),$$

where bold font is being used temporarily to denote mathematical objects in the semantics to distinguish them from symbols in the syntax. To logicize this belief, first, the syntax has to be specified. For this, an alphabet is defined. Thus the type constructors are specifed and each basic mathematical object in the definition is denoted by a variable or constant in the alphabet. For example, let σ be the type of elements in X, τ the type of elements in Y, and f a constant that denotes $\mathbf{f}$. Then f is a constant having signature $\sigma \to \tau$. The alphabet has to be rich enough so that each basic mathematical object in $\mathbf{t}$ is denoted by a symbol in the alphabet. And, of course, the alphabet must be rich enough so that this can be done for every belief in the agent's belief base.

Now the intended pointed interpretation for the alphabet can be defined. An interpretation is a quadruple consisting of a set W of worlds, a set $\{R_i\}_{i=1}^m$ of accessibility relations on W, a set of domains $\{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}$, and a valuation V that specifies the denotation (that is, meaning) for each constant in the alphabet in each world. (See Definition B.2.5.) A

pointed interpretation (I, w) is an interpretation I with a distinguished world $w \in W$ that is the point. Intuitively, w is the actual world and the worlds in W that are accessible to w are 'nearby' worlds where the denotations of the symbols in the alphabet are possibly different from their denotations in w. A variable assignment ν is a mapping from each variable of type α to an element of the domain $\mathcal{D}_\alpha$. With these ingredients, the denotation $\mathcal{V}(t, I, w, \nu)$ of a term t of type α in the language of terms given by the alphabet is defined in a natural way to be an element of $\mathcal{D}_\alpha$. (See Definition B.2.10. For closed terms, the case of interest in the following two paragraphs, the variable assignment ν can be ignored.)

Now the logicization process can be completed by writing down suitable formulas (or, more generally, biterms) in a theory in the logic that logicize the beliefs, that is, give the logical representations of the beliefs. For example, concentrating on the belief $\mathbf{f}$, there may be a formula in this theory that has the form

$$\Box\forall x.((f\ x) = t),$$

where $\Box$ is some sequence of modalities and t is a term of type τ. Equivalently, by Proposition B.2.24, this formula could take the form

$$\Box(f = \lambda x.t).$$

A crucial requirement is that, if (I, w) is the intended pointed interpretation, then $\Box\forall x.(f(x) = t)$ is valid at w in I, that is,

$$\mathcal{V}(\Box\forall x.(f(x) = t), I, w) = \mathsf{T}.$$

(Since $\Box\forall x.(f(x) = t)$ is assumed to be closed, reference in $\mathcal{V}$ to any variable assignment can be dropped. See Proposition B.2.14 and the immediately following notation.) As an example, suppose that $\Box$ is $\bullet\boldsymbol{B}_i\boldsymbol{B}_j$, where $\bullet$ is the 'last time' temporal modality, $\boldsymbol{B}_i$ is the belief modality for agent i, and $\boldsymbol{B}_j$ is the belief modality for agent j. Thus the formula $\Box\forall x.((f\ x) = t)$ is

$$\bullet\boldsymbol{B}_i\boldsymbol{B}_j\forall x.((f\ x) = t).$$

The modalities capture the notion that, at the last time, agent i believed that agent j believed that $\forall x.((f\ x) = t)$ held. The intention is that the denotation of $\bullet\boldsymbol{B}_i\boldsymbol{B}_j f$ with respect to the intended pointed interpretation (I, w) be $\mathbf{f}$, that is,

$$\mathcal{V}(\bullet\boldsymbol{B}_i\boldsymbol{B}_j f, I, w) = \mathbf{f}.$$

Note that, since $\mathcal{V}(\bullet\boldsymbol{B}_i\boldsymbol{B}_j(f = \lambda x.t), I, w) = \mathsf{T}$, it follows, by Proposition B.2.22, that $\mathcal{V}(\bullet\boldsymbol{B}_i\boldsymbol{B}_j f = \bullet\boldsymbol{B}_i\boldsymbol{B}_j\lambda x.t, I, w) = \mathsf{T}$, and hence, by Proposition B.2.10, that

$$\mathcal{V}(\bullet\boldsymbol{B}_i\boldsymbol{B}_j f, I, w) = \mathcal{V}(\bullet\boldsymbol{B}_i\boldsymbol{B}_j\lambda x.t, I, w).$$

There may be another agent k and a formula

$$\bullet\boldsymbol{B}_i\boldsymbol{B}_k\forall x.((f\ x) = s)$$

that gives the logical representation of a belief of agent i. Thus agent i has beliefs, the denotations of $\bullet\boldsymbol{B}_i\boldsymbol{B}_j f$ and $\bullet\boldsymbol{B}_i\boldsymbol{B}_k f$, about the beliefs of two other agents j and k, and

both these beliefs concern the constant f. Generally, these two beliefs are different, that is, $\mathcal{V}(\bullet \boldsymbol{B}_i \boldsymbol{B}_j f, I, w) \neq \mathcal{V}(\bullet \boldsymbol{B}_i \boldsymbol{B}_k f, I, w)$. For example, suppose that agent i is a poker-playing agent and agents j and k are opponents (human or otherwise). The intended meaning of the constant f is that, depending on the information available to a player, the player should either fold or not fold a hand. Thus $\mathcal{V}(\bullet \boldsymbol{B}_i \boldsymbol{B}_j f, I, w)$ is the belief of agent i at the last time that agent j has a certain belief that it uses to fold or not. Similarly, for agent k and the belief $\mathcal{V}(\bullet \boldsymbol{B}_i \boldsymbol{B}_k f, I, w)$. Typically, agent i will believe agents j and k use different beliefs to decide whether to fold or not. Note also that $\mathcal{V}(\bullet \boldsymbol{B}_i \boldsymbol{B}_j f, I, w)$ will generally be different from $\mathcal{V}(\bullet \boldsymbol{B}_j f, I, w)$, a belief of agent j. Similarly, with respect to a pointed interpretation (I, w), the beliefs that are the denotations of $\bullet \boldsymbol{B}_i \boldsymbol{B}_k f$, $\bullet \boldsymbol{B}_k f$, $\boldsymbol{B}_i f$, $\boldsymbol{B}_j f$, $\boldsymbol{B}_k f$, and f, for example, will generally all be different from one another.

This logicization provides an additional perspective about beliefs: A belief held by an agent i is often the denotation with respect to the intended pointed interpretation of a term of the form $\Box f$, where f is a constant having signature of the form $\sigma \to \tau$ and the leftmost modality amongst the belief (or knowledge) modalities in $\Box$ is $\boldsymbol{B}_i$.

The term 'belief' has two meanings in artificial intelligence: in robotics and vision, a belief is generally a probability distribution; in logical artificial intelligence, a belief is a formula. The preceding analysis of the logicization of beliefs comes down heavily in favour of the robotics and vision view (generalized appropriately). Beliefs are in the semantics, not the syntax. A belief is a function, not a formula!

In this spirit, here is some terminology that is intended to maintain a clear distinction between the syntax and the semantics of beliefs. For a particular agent in some application, the *belief theory* is the theory that results from representing each belief of the agent in the logic. Thus the belief theory is the theory that results from the logicization of the belief base of the agent. Each assumption that is a formula in a belief theory is called a *belief formula*. Similarly, each assumption that is a biterm in a belief theory is called a *belief biterm*.

Note carefully that, in practice, one usually works only with the theory as (non-empirical) beliefs themselves are never materialized. So, for belief $\mathbf{f}$, for example, all that is needed is the formula $\bullet \boldsymbol{B}_i \boldsymbol{B}_j \forall x.((f\ x) = t)$, not $\mathbf{f}$ itself. In general, the theory must be sufficiently detailed so that any information about beliefs needed to choose actions can be obtained by reasoning about the theory. However, for *empirical* beliefs, both the belief and the corresponding belief formula(s) *are* materialized. Empirical beliefs change over time and one needs the actual belief to perform filtering. A simple computation is then needed to produce the corresponding belief formula(s) that are needed for reasoning.

To do: Give the details of the logicization process.

5.3 Computation Examples

This section contains several examples to illustrate computation, which is described in detail in Section B.3.1.

Example 5.3.1. Consider the following polymorphic definitions of the constants *append*, *permute*, *delete*, and *sorted* which have been written in the relational style of logic pro-

gramming.

$$
\begin{array}{l}
append\ :\ List\ a \times List\ a \times List\ a \to Bool \\
(append\ (u, v, w)) = \\
\quad ((u = []) \wedge (v = w)) \vee \exists r.\exists x.\exists y.((u = r \sharp x) \wedge (w = r \sharp y) \wedge (append\ (x, v, y)))
\end{array}
$$

$$
\begin{array}{l}
permute\ :\ List\ a \times List\ a \to Bool \\
(permute\ ([], x)) = x = [] \\
(permute\ (x \sharp y, w)) = \\
\quad \exists u.\exists v.\exists z.((w = u \sharp v) \wedge (delete\ (u, x \sharp y, z)) \wedge (permute\ (z, v)))
\end{array}
$$

$$
\begin{array}{l}
delete\ :\ a \times List\ a \times List\ a \to Bool \\
(delete\ (x, [], y)) = \bot \\
(delete\ (x, y \sharp z, w)) = \\
\quad\quad ((x = y) \wedge (w = z)) \vee \exists v.((w = y \sharp v) \wedge (delete\ (x, z, v)))
\end{array}
$$

$$
\begin{array}{l}
sorted\ :\ List\ Int \to Bool \\
(sorted\ []) = \top \\
(sorted\ x \sharp y) = \\
\quad if\ y = []\ then\ \top\ else\ \exists u.\exists v.((y = u \sharp v) \wedge (x \leq u) \wedge (sorted\ y)).
\end{array}
$$

The intended meaning of *append* is that it is true iff its third argument is the concatenation of its first two arguments. The intended meaning of *permute* is that it is true iff its second argument is a permutation of its first argument. The intended meaning of *delete* is that it is true iff its third argument is the result of deleting its first argument from its second argument. The intended meaning of *sorted* is that it is true iff its argument is an increasingly ordered list of integers.

The notable feature of the above definitions is the presence of existential quantifiers in the bodies of the statements, so not surprisingly the key statement that makes all this work is concerned with the existential quantifier. To motivate this, consider the computation in Figure 5.1 that results from the goal $(append\ (1 \sharp [], 2 \sharp [], x))$. At one point in the computation, the following term is reached:

$$\exists r'.\exists x'.\exists y'.((1 = r') \wedge ([] = x') \wedge (x = r' \sharp y') \wedge (append\ (x', 2 \sharp [], y'))).$$

An obviously desirable simplification that can be made to this term is to eliminate the local variable r' since there is a 'value' (that is, 1) for it. This leads to the term

$$\exists x'.\exists y'.(([] = x') \wedge (x = 1 \sharp y') \wedge (append\ (x', 2 \sharp [], y'))).$$

Similarly, one can eliminate x' to obtain

$$\exists y'.((x = 1 \sharp y') \wedge (append\ ([], 2 \sharp [], y'))).$$

After some more computation, the answer $x = 1 \sharp 2 \sharp []$ results. Now the statements that make all this possible are

$$\exists x_1. \cdots \exists x_n.(\boldsymbol{x} \wedge (x_i = \boldsymbol{u}) \wedge \boldsymbol{y}) = \exists x_1. \cdots \exists x_{i-1}. \exists x_{i+1}. \cdots \exists x_n.(\boldsymbol{x}\{x_i/\boldsymbol{u}\} \wedge \boldsymbol{y}\{x_i/\boldsymbol{u}\})$$
$$\exists x_1. \cdots \exists x_n.(\boldsymbol{x} \wedge (\boldsymbol{u} = x_i) \wedge \boldsymbol{y}) = \exists x_1. \cdots \exists x_{i-1}. \exists x_{i+1}. \cdots \exists x_n.(\boldsymbol{x}\{x_i/\boldsymbol{u}\} \wedge \boldsymbol{y}\{x_i/\boldsymbol{u}\}),$$

which come from the definition of $\Sigma : (a \to \mathit{Bool}) \to \mathit{Bool}$ in the standard equality theory in Section B.3.1 and have λ-abstractions in their heads.

$$\underline{(\mathit{append}\ (1 \sharp [], 2 \sharp [], x))}$$
$$(\underline{(1 \sharp [] = [])} \wedge (2 \sharp [] = x)) \vee \exists r'. \exists x'. \exists y'. ((1 \sharp [] = r' \sharp x') \wedge (x = r' \sharp y') \wedge (\mathit{append}\ (x', 2 \sharp [], y')))$$
$$\underline{(\bot \wedge (2 \sharp [] = x))} \vee \exists r'. \exists x'. \exists y'. ((1 \sharp [] = r' \sharp x') \wedge (x = r' \sharp y') \wedge (\mathit{append}\ (x', 2 \sharp [], y')))$$
$$\underline{\bot \vee \exists r'. \exists x'. \exists y'. ((1 \sharp [] = r' \sharp x') \wedge (x = r' \sharp y') \wedge (\mathit{append}\ (x', 2 \sharp [], y')))}$$
$$\exists r'. \exists x'. \exists y'. (\underline{(1 \sharp [] = r' \sharp x')} \wedge (x = r' \sharp y') \wedge (\mathit{append}\ (x', 2 \sharp [], y')))$$
$$\underline{\exists r'. \exists x'. \exists y'. ((1 = r') \wedge ([] = x') \wedge (x = r' \sharp y') \wedge (\mathit{append}\ (x', 2 \sharp [], y')))}$$
$$\underline{\exists x'. \exists y'. (([] = x') \wedge (x = 1 \sharp y') \wedge (\mathit{append}\ (x', 2 \sharp [], y')))}$$
$$\exists y'. ((x = 1 \sharp y') \wedge \underline{(\mathit{append}\ ([], 2 \sharp [], y'))})$$
$$\vdots$$
$$\underline{\exists y'. ((x = 1 \sharp y') \wedge (y' = 2 \sharp []))}$$
$$x = 1 \sharp 2 \sharp []$$

Figure 5.1: Computation of $(\mathit{append}\ (1 \sharp [], 2 \sharp [], x))$

This example illustrates how the definitions in the standard equality theory allow the traditional functional programming style to be extended to encompass the relational style of logic programming. The definitions of predicates look a little different from the way one would write them in, say, Prolog. A mechanical translation of a Prolog definition into one that runs in this style of programming simply involves using the completion of the Prolog definition. The definition here of *append* is essentially the completion of the Prolog version of *append*. Alternatively, one can specialize the completion to the [] and $\sharp$ cases, as has been done here for the definitions of *permute*, *delete*, and *sorted*. One procedural difference of note is that Prolog's method of returning answers one at a time via backtracking is replaced here by returning all answers together as a disjunction (or a set). Thus the computation in Figure 5.2 from the goal

$$(\mathit{append}\ (x, y, 1 \sharp 2 \sharp []))$$

has the answer

$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee ((x = 1 \sharp 2 \sharp []) \wedge (y = [])).$$

The style of programming typified by this example is called *programming with abstractions.*

Example 5.3.2. Consider again the constants *sometime* and *upto_sometime* that were introduced in Section 5.1 and that have the following definitions.

$$sometime : Bool \times Nat \to Bool$$
$$(sometime\ (\boldsymbol{\varphi}, n)) = if\ n = 0\ then\ \boldsymbol{\varphi}\ else\ \bigcirc(sometime\ (\boldsymbol{\varphi}, n-1))$$

$$upto\_sometime : Bool \times Nat \to Bool$$
$$(upto\_sometime\ (\boldsymbol{\varphi}, n)) =$$
$$\quad if\ n = 0\ then\ \boldsymbol{\varphi}\ else\ ((sometime\ (\boldsymbol{\varphi}, n)) \wedge (upto\_sometime\ (\boldsymbol{\varphi}, n-1))).$$

These schemes are taken to be global assumptions of the theory.

Now let φ be some biterm. Then Figure 5.3 gives a computation of $(upto\_sometime\ (\varphi, 2))$. The computation uses the leftmost selection rule and has the redexes underlined. It follows from the computation that

$$(upto\_sometime\ (\varphi, 2)) = \bigcirc^{\leq 2}\varphi$$

is a consequence of the theory.

Example 5.3.3. Consider a belief theory for an agent that contains the definition

$$\boldsymbol{B}\ \forall x.((f\ x) = if\ x = A\ then\ 42\ else\ if\ x = B\ then\ 21\ else\ if\ x = C\ then\ 42\ else\ 0),$$

where $A, B, C : \sigma$, $f : \sigma \to Nat$ and $\boldsymbol{B}$ is the belief modality for the agent. With such a definition, it is straightforward to compute in the 'forward' direction. Thus $(f\ B)$ can be computed in the obvious way to produce the answer 21 and the result $\boldsymbol{B}((f\ B) = 21)$.

Less obviously, the definition can be used to compute in the 'reverse' direction. For example, consider the computation of $\{x \mid (f\ x) = 42\}$ in Figure 5.4, which produces the answer $\{A, C\}$. This computation makes essential use of the equations

$$(\boldsymbol{w}\ \ if\ \boldsymbol{x} = \boldsymbol{t}\ then\ u\ else\ v) = if\ \boldsymbol{x} = \boldsymbol{t}\ then\ (\boldsymbol{w}\{\boldsymbol{x}/\boldsymbol{t}\}\ u)\ else\ (\boldsymbol{w}\ v)$$
$$\quad \% \text{ where } \boldsymbol{x} \text{ is a variable.}$$
$$(if\ \boldsymbol{x} = \boldsymbol{t}\ then\ u\ else\ v\ \ \boldsymbol{w}) = if\ \boldsymbol{x} = \boldsymbol{t}\ then\ (u\ \boldsymbol{w}\{\boldsymbol{x}/\boldsymbol{t}\})\ else\ (v\ \boldsymbol{w})$$
$$\quad \% \text{ where } \boldsymbol{x} \text{ is a variable.}$$

from the standard equality theory.

Example 5.3.4. This example illustrates how typical database queries can be answered. Consider the definition

$$\boldsymbol{B}\ \forall x.((f\ x) = if\ x = (A, Z)\ then\ (42, 11)\ else\ if\ x = (B, X)\ then\ (21, 7)\ else$$
$$if\ x = (C, Y)\ then\ (42, 17)\ else\ (0, 0)),$$

$$\underline{(append\ (x, y, 1 \sharp 2 \sharp []))}$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee \exists r'.\exists x'.\exists y'.((x = r' \sharp x') \wedge \underline{(1 \sharp 2 \sharp [] = r' \sharp y')} \wedge (append\ (x', y, y')))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee \underline{\exists r'.\exists x'.\exists y'.((x = r' \sharp x') \wedge (1 = r') \wedge (2 \sharp [] = y') \wedge (append\ (x', y, y')))}$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee \underline{\exists x'.\exists y'.((x = 1 \sharp x') \wedge (2 \sharp [] = y') \wedge (append\ (x', y, y')))}$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee \exists x'.((x = 1 \sharp x') \wedge \underline{(append\ (x', y, 2 \sharp []))})$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee \exists x'.\underline{((x = 1 \sharp x') \wedge (((x' = []) \wedge (y = 2 \sharp [])) \vee}$$
$$\underline{\exists r''.\exists x''.\exists y''.((x' = r'' \sharp x'') \wedge (2 \sharp [] = r'' \sharp y'') \wedge (append\ (x'', y, y''))))}$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee \underline{\exists x'.(((x = 1 \sharp x') \wedge (x' = []) \wedge (y = 2 \sharp [])) \vee}$$
$$\underline{((x = 1 \sharp x') \wedge \exists r''.\exists x''.\exists y''.((x' = r'' \sharp x'') \wedge (2 \sharp [] = r'' \sharp y'') \wedge (append\ (x'', y, y''))))}$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee \underline{\exists x'.((x = 1 \sharp x') \wedge (x' = []) \wedge (y = 2 \sharp []))} \vee$$
$$\exists x'.((x = 1 \sharp x') \wedge \exists r''.\exists x''.\exists y''.((x' = r'' \sharp x'') \wedge (2 \sharp [] = r'' \sharp y'') \wedge (append\ (x'', y, y''))))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee$$
$$\exists x'.((x = 1 \sharp x') \wedge \exists r''.\exists x''.\exists y''.((x' = r'' \sharp x'') \wedge \underline{(2 \sharp [] = r'' \sharp y'')} \wedge (append\ (x'', y, y''))))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee$$
$$\exists x'.((x = 1 \sharp x') \wedge \underline{\exists r''.\exists x''.\exists y''.((x' = r'' \sharp x'') \wedge (2 = r'') \wedge ([] = y'') \wedge (append\ (x'', y, y'')))})$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee$$
$$\exists x'.((x = 1 \sharp x') \wedge \underline{\exists x''.\exists y''.((x' = 2 \sharp x'') \wedge ([] = y'') \wedge (append\ (x'', y, y'')))})$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee$$
$$\exists x'.((x = 1 \sharp x') \wedge \exists x''.((x' = 2 \sharp x'') \wedge \underline{(append\ (x'', y, []))}))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee \exists x'.((x = 1 \sharp x') \wedge \exists x''.((x' = 2 \sharp x'') \wedge$$
$$(((x'' = []) \wedge (y = [])) \vee \exists r'''.\exists x'''.\exists y'''.((x'' = r''' \sharp x''') \wedge \underline{([] = r''' \sharp y''')} \wedge (append\ (x''', y, y'''))))))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee \exists x'.((x = 1 \sharp x') \wedge \exists x''.((x' = 2 \sharp x'') \wedge$$
$$(((x'' = []) \wedge (y = [])) \vee \exists r'''.\exists x'''.\exists y'''.(\underline{(x'' = r''' \sharp x''') \wedge \bot} \wedge (append\ (x''', y, y'''))))))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee \exists x'.((x = 1 \sharp x') \wedge \exists x''.((x' = 2 \sharp x'') \wedge$$
$$(((x'' = []) \wedge (y = [])) \vee \exists r'''.\exists x'''.\exists y'''.\underline{(\bot \wedge (append\ (x''', y, y''')))})))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee$$
$$\exists x'.((x = 1 \sharp x') \wedge \exists x''.((x' = 2 \sharp x'') \wedge (((x'' = []) \wedge (y = [])) \vee \underline{\exists r'''.\exists x'''.\exists y'''.\bot})))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee$$
$$\exists x'.((x = 1 \sharp x') \wedge \exists x''.((x' = 2 \sharp x'') \wedge \underline{(((x'' = []) \wedge (y = [])) \vee \bot)}))$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee$$
$$\exists x'.((x = 1 \sharp x') \wedge \underline{\exists x''.((x' = 2 \sharp x'') \wedge (x'' = []) \wedge (y = []))})$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee$$
$$\underline{\exists x'.((x = 1 \sharp x') \wedge (x' = 2 \sharp []) \wedge (y = []))}$$
$$((x = []) \wedge (y = 1 \sharp 2 \sharp [])) \vee ((x = 1 \sharp []) \wedge (y = 2 \sharp [])) \vee ((x = 1 \sharp 2 \sharp []) \wedge (y = []))$$

Figure 5.2: Computation of $(append\ (x, y, 1 \sharp 2 \sharp []))$

$$\underline{(\mathit{upto\_sometime}\ (\varphi, 2))}$$
$$\mathit{if}\ \underline{2 = 0}\ \mathit{then}\ \varphi\ \mathit{else}\ ((\mathit{sometime}\ (\varphi, 2)) \wedge (\mathit{upto\_sometime}\ (\varphi, 2-1)))$$
$$\underline{\mathit{if}\ \bot\ \mathit{then}\ \varphi\ \mathit{else}\ ((\mathit{sometime}\ (\varphi, 2)) \wedge (\mathit{upto\_sometime}\ (\varphi, 2-1)))}$$
$$\underline{(\mathit{sometime}\ (\varphi, 2))} \wedge (\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$(\mathit{if}\ \underline{2 = 0}\ \mathit{then}\ \varphi\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\varphi, 2-1))) \wedge (\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$(\underline{\mathit{if}\ \bot\ \mathit{then}\ \varphi\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\varphi, 2-1))}) \wedge (\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\bigcirc(\underline{\mathit{sometime}\ (\varphi, 2-1)}) \wedge (\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\bigcirc\,(\mathit{if}\ \underline{(2-1)} = 0\ \mathit{then}\ \varphi\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\varphi, (2-1)-1))) \wedge$$
$$(\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\bigcirc\,(\mathit{if}\ \underline{1 = 0}\ \mathit{then}\ \varphi\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\varphi, (2-1)-1))) \wedge$$
$$(\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\bigcirc\,(\underline{\mathit{if}\ \bot\ \mathit{then}\ \varphi\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\varphi, (2-1)-1))}) \wedge$$
$$(\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\bigcirc\bigcirc(\underline{\mathit{sometime}\ (\varphi, (2-1)-1)}) \wedge (\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\bigcirc\bigcirc\,(\mathit{if}\ \underline{(2-1)} - 1 = 0\ \mathit{then}\ \varphi\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\varphi, ((2-1)-1)-1))) \wedge$$
$$(\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\vdots$$
$$\bigcirc\bigcirc\,(\mathit{if}\ \underline{0 = 0}\ \mathit{then}\ \varphi\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\varphi, ((2-1)-1)-1))) \wedge$$
$$(\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\bigcirc\bigcirc\,(\underline{\mathit{if}\ \top\ \mathit{then}\ \varphi\ \mathit{else}\ \bigcirc(\mathit{sometime}\ (\varphi, ((2-1)-1)-1))}) \wedge$$
$$(\mathit{upto\_sometime}\ (\varphi, 2-1))$$
$$\bigcirc\bigcirc\varphi \wedge (\underline{\mathit{upto\_sometime}\ (\varphi, 2-1)})$$
$$\vdots$$
$$\bigcirc\bigcirc\varphi \wedge \bigcirc\varphi \wedge (\underline{\mathit{upto\_sometime}\ (\varphi, (2-1)-1)})$$
$$\vdots$$
$$\bigcirc\bigcirc\varphi \wedge \bigcirc\varphi \wedge \varphi$$

Figure 5.3: Computation of $(\mathit{upto\_sometime}\ (\varphi, 2))$

$\{x \mid \underline{(f\ x)} = 42\}$
$\{x \mid (\underline{(=\ \ \textit{if}\ x = A\ \textit{then}\ 42\ \textit{else if}\ x = B\ \textit{then}\ 21\ \textit{else if}\ x = C\ \textit{then}\ 42\ \textit{else}\ 0)}\ 42)\}$
$\{x \mid \underline{(\textit{if}\ x = A\ \textit{then}\ (=\ 42)\ \textit{else}\ (=\ \ \textit{if}\ x = B\ \textit{then}\ 21\ \textit{else if}\ x = C\ \textit{then}\ 42\ \textit{else}\ 0)\ 42)}\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \underline{(42 = 42)}\ \textit{else}\ ((=\ \ \textit{if}\ x = B\ \textit{then}\ 21\ \textit{else if}\ x = C\ \textit{then}\ 42\ \textit{else}\ 0)\ 42)\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else}\ \underline{((=\ \ \textit{if}\ x = B\ \textit{then}\ 21\ \textit{else if}\ x = C\ \textit{then}\ 42\ \textit{else}\ 0)}\ 42)\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else}\ \underline{(\textit{if}\ x = B\ \textit{then}\ (=\ 21)\ \textit{else}\ (=\ \ \textit{if}\ x = C\ \textit{then}\ 42\ \textit{else}\ 0)\ 42)}\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else if}\ x = B\ \textit{then}\ \underline{(21 = 42)}\ \textit{else}\ ((=\ \ \textit{if}\ x = C\ \textit{then}\ 42\ \textit{else}\ 0)\ 42)\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else if}\ x = B\ \textit{then}\ \bot\ \textit{else}\ \underline{((=\ \ \textit{if}\ x = C\ \textit{then}\ 42\ \textit{else}\ 0)}\ 42)\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else if}\ x = B\ \textit{then}\ \bot\ \textit{else}\ \underline{(\textit{if}\ x = C\ \textit{then}\ (=\ 42)\ \textit{else}\ (=\ 0)\ \ 42)}\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else if}\ x = B\ \textit{then}\ \bot\ \textit{else if}\ x = C\ \textit{then}\ \underline{(42 = 42)}\ \textit{else}\ (0 = 42)\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else if}\ x = B\ \textit{then}\ \bot\ \textit{else if}\ x = C\ \textit{then}\ \top\ \textit{else}\ \underline{(0 = 42)}\}$
$\{x \mid \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else if}\ x = B\ \textit{then}\ \bot\ \textit{else if}\ x = C\ \textit{then}\ \top\ \textit{else}\ \bot\}$

Figure 5.4: Computation of rank 0 using $\boldsymbol{B}$ of $\{x \mid (f\ x) = 42\}$

where $A, B, C : \sigma$, $X, Y, Z : \tau$ and $f : \sigma \times \tau \to Nat \times Nat$.

Figure 5.5 shows the computation using $\boldsymbol{B}$ of

$$\exists y.((y = (f\ (C, v))) \wedge ((proj_1\ y) = 42)),$$

where $proj_1$ projects onto the first component of a pair of natural numbers. The result of the computation is $\boldsymbol{B}(\exists y.((y = (f\ (C, v))) \wedge ((proj_1\ y) = 42)) = (v = Y))$ and the answer is $v = Y$.

Figure 5.6 shows the computation using $\boldsymbol{B}$ of

$$\{y \mid \exists x.((y = (f\ x)) \wedge ((proj_1\ y) = 42))\}.$$

The result of the computation is $\boldsymbol{B}(\{y \mid \exists x.((y = (f\ x)) \wedge ((proj_1\ y) = 42))\} = \{(42, 11), (42, 17)\})$ and the answer is $\{(42, 11), (42, 17)\}$.

Example 5.3.5. Consider the following theory that contains a record of current and past statistics on the price of a commodity.

$prices : Density\ Real$
$prices = (gaussian\ 400\ 20)$ (5.3.1)
$\bullet(prices = (gaussian\ 360\ 25))$ (5.3.2)
$\bullet^2(prices = \lambda x.if\ x = 300\ then\ 0.7\ else\ if\ x = 310\ then\ 0.3\ else\ 0)$ (5.3.3)
$\bullet^3(prices = (gaussian\ 330\ 10))$ (5.3.4)
$\bullet^4\blacksquare(prices = \lambda x.if\ x = 280\ then\ 1\ else\ 0)$ (5.3.5)

$gaussian : Real \to Real \to Density\ Real$

$\underline{\exists y.((y = (f\ (C, v))) \wedge ((proj_1\ y) = 42))}$
$(proj_1\ \underline{(f\ (C, v))}) = 42$
$\underline{(proj_1\ \mathit{if}\ (C, v) = (A, Z)\ \mathit{then}\ (42, 11)\ \mathit{else}\ \mathit{if}\ (C, v) = (B, X)\ \mathit{then}\ (21, 7)}$
$\underline{\mathit{else}\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0))}\ =\ 42$
$(\underline{(=\ \ \mathit{if}\ (C, v) = (A, Z)\ \mathit{then}\ (proj_1\ (42, 11))\ \mathit{else}\ (proj_1\ \mathit{if}\ (C, v) = (B, X)\ \mathit{then}\ (21, 7)}$
$\underline{\mathit{else}\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0)))}\ 42)$
$\underline{(\mathit{if}\ (C, v) = (A, Z)\ \mathit{then}\ (=\ (proj_1\ (42, 11)))\ \mathit{else}\ (=\ (proj_1\ \mathit{if}\ (C, v) = (B, X)\ \mathit{then}\ (21, 7)}$
$\underline{\mathit{else}\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0)))\ 42)}$
$\mathit{if}\ \underline{(C, v) = (A, Z)}\ \mathit{then}\ ((proj_1\ (42, 11)) = 42)\ \mathit{else}\ ((proj_1\ \mathit{if}\ (C, v) = (B, X)\ \mathit{then}\ (21, 7)$
$\mathit{else}\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0)) = 42)$
$\mathit{if}\ \underline{(C = A) \wedge (v = Z)}\ \mathit{then}\ ((proj_1\ (42, 11)) = 42)\ \mathit{else}\ ((proj_1\ \mathit{if}\ (C, v) = (B, X)\ \mathit{then}\ (21, 7)$
$\mathit{else}\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0)) = 42)$
$\mathit{if}\ \underline{\bot \wedge (v = Z)}\ \mathit{then}\ ((proj_1\ (42, 11)) = 42)\ \mathit{else}\ ((proj_1\ \mathit{if}\ (C, v) = (B, X)\ \mathit{then}\ (21, 7)$
$\mathit{else}\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0)) = 42)$
$\underline{\mathit{if}\ \bot\ \mathit{then}\ ((proj_1\ (42, 11)) = 42)\ \mathit{else}\ ((proj_1\ \mathit{if}\ (C, v) = (B, X)\ \mathit{then}\ (21, 7)}$
$\underline{\mathit{else}\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0)) = 42)}$
$\underline{(proj_1\ \mathit{if}\ (C, v) = (B, X)\ \mathit{then}\ (21, 7)\ \mathit{else}\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0))} = 42$
$\vdots$
$\underline{(proj_1\ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (42, 17)\ \mathit{else}\ (0, 0))} = 42$
$(\underline{(=\ \ \mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (proj_1\ (42, 17))\ \mathit{else}\ (proj_1\ (0, 0)))}\ 42)$
$\underline{(\mathit{if}\ (C, v) = (C, Y)\ \mathit{then}\ (=\ (proj_1\ (42, 17)))\ \mathit{else}\ (=\ (proj_1\ (0, 0)))\ 42)}$
$\mathit{if}\ \underline{(C, v) = (C, Y)}\ \mathit{then}\ ((proj_1\ (42, 17)) = 42)\ \mathit{else}\ ((proj_1\ (0, 0)) = 42)$
$\mathit{if}\ \underline{(C = C) \wedge (v = Y)}\ \mathit{then}\ ((proj_1\ (42, 17)) = 42)\ \mathit{else}\ ((proj_1\ (0, 0)) = 42)$
$\mathit{if}\ \underline{\top \wedge (v = Y)}\ \mathit{then}\ ((proj_1\ (42, 17)) = 42)\ \mathit{else}\ ((proj_1\ (0, 0)) = 42)$
$\mathit{if}\ (v = Y)\ \mathit{then}\ (\underline{(proj_1\ (42, 17))} = 42)\ \mathit{else}\ ((proj_1\ (0, 0)) = 42)$
$\mathit{if}\ (v = Y)\ \mathit{then}\ \underline{(42 = 42)}\ \mathit{else}\ ((proj_1\ (0, 0)) = 42)$
$\mathit{if}\ (v = Y)\ \mathit{then}\ \top\ \mathit{else}\ (\underline{(proj_1\ (0, 0))} = 42)$
$\mathit{if}\ (v = Y)\ \mathit{then}\ \top\ \mathit{else}\ \underline{(0 = 42)}$
$\mathit{if}\ (v = Y)\ \mathit{then}\ \top\ \mathit{else}\ \bot$

Figure 5.5: Computation of rank 0 using $\boldsymbol{B}$ of $\exists y.((y = (f\ (C, v))) \wedge ((proj_1\ y) = 42))$

$$
\begin{array}{l}
\{y \mid \exists x.((y = \underline{(f\ x)}) \wedge ((proj_1\ y) = 42))\} \\
\{y \mid \exists x.(\underline{(y = \varphi) \wedge ((proj_1\ y) = 42)})\} \\
\{y \mid \exists x.((y = \varphi) \wedge (\underline{(proj_1\ \varphi)} = 42))\} \\
\quad \vdots \\
\{y \mid \exists x.(\underline{(y = \varphi) \wedge (if\ x = (A, Z)\ then\ \top\ else\ if\ x = (B, X)\ then\ \bot} \\
\qquad\qquad \underline{else\ if\ x = (C, Y)\ then\ \top\ else\ \bot)})\} \\
\{y \mid \underline{\exists x.(if\ ((y = \varphi) \wedge (x = (A, Z)))\ then\ \top\ else} \\
\quad \underline{if\ ((y = \varphi) \wedge (if\ x = (B, X)\ then\ \bot\ else\ if\ x = (C, Y)\ then\ \top\ else\ \bot)))}\} \\
\{y \mid if\ \underline{\exists x.((y = \varphi) \wedge (x = (A, Z)))}\ then\ \top\ else \\
\quad \exists x.(if\ ((y = \varphi) \wedge (if\ x = (B, X)\ then\ \bot\ else\ if\ x = (C, Y)\ then\ \top\ else\ \bot)))\} \\
\quad \vdots \\
\{y \mid if\ y = (42, 11)\ then\ \top\ else\ if\ \exists x.(\underline{(y = \varphi) \wedge (if\ x = (B, X)\ then\ \bot} \\
\qquad\qquad \underline{else\ if\ x = (C, Y)\ then\ \top\ else\ \bot)})\} \\
\quad \vdots \\
\{y \mid if\ y = (42, 11)\ then\ \top\ else\ if\ y = (21, 7)\ then\ \bot\ else\ if\ y = (42, 17)\ then\ \top\ else\ \bot\}
\end{array}
$$

Figure 5.6: Computation of rank 0 using $\boldsymbol{B}$ of $\{y \mid \exists x.((y = (f\ x)) \wedge ((proj_1\ y) = 42))\}$. Here, $\varphi \triangleq if\ x = (A, Z)\ then\ (42, 11)\ else\ if\ x = (B, X)\ then\ (21, 7)\ else\ if\ x = (C, Y)\ then\ (42, 17)\ else\ (0, 0)$

$$(gaussian\ u\ s) = \lambda x.\frac{1}{s\sqrt{2\pi}}e^{-\frac{(x-u)^2}{2s^2}} \tag{5.3.6}$$

$$mean : (Density\ a) \rightarrow Real$$

$$(mean\ (gaussian\ u\ s)) = u \tag{5.3.7}$$

$$(mean\ \lambda x.0) = 0 \tag{5.3.8}$$

$$(mean\ \lambda x.if\ x = u\ then\ y\ else\ \mathbf{w}) = y \times u + (mean\ \lambda x.\mathbf{w}) \tag{5.3.9}$$

$$\blacklozenge x = (x \vee \bullet\blacklozenge x). \tag{5.3.10}$$

Here, *mean* and *gaussian* are rigid constants whereas *prices* is not. An example query one might want to ask is

$$\blacklozenge((mean\ prices) < (mean\ \bullet prices)).$$

In other words, is there a period in the past where mean prices fell? The computation should return the answer $\top$ in this case. (The $\mathbf{w}$ in the third equation for *mean* is a syntactical variable; this is explained in more detail in Appendix B.1.)

Figure 5.7 shows the computation of $\blacklozenge((mean\ prices) < (mean\ \bullet prices))$ using the

program of Example 5.3.5. (The labels O1, O2, and M refer to particular equations in the standard equality theory in Section B.3.1.) Among other things, the computation shows

1. how redexes made up of non-rigid terms can only be rewritten using definitions with the correct modal context;
2. how global assumptions can be used inside any modal context;
3. how probability densities can be manipulated using higher-order functions; and
4. how syntactical variables are used to process lambda abstractions.

5.4 Proof Examples

Here are several examples to illustrate proof, which is described in detail in Section B.3.2.

Example 5.4.1. Suppose a theory consists of just the local assumption $\forall x.\Box_i\varphi$, where φ is some biterm. Then the proof in Figure 5.8 shows that $\Box_i\forall x.\varphi$ is a consequence of the theory.

An explanation of this proof is as follows. Item 1 is the negation of the biterm to be proved; 2 is from 1 by a possibility rule; 3 is from 2 by an existential rule; 4 is a local assumption; 5 is from 4 by a universal rule; 6 is from 5 by a necessity rule; now the branch closes by 3 and 6.

In effect, this proof establishes the Barcan biterm

$$\forall x.\Box_i\varphi \longrightarrow \Box_i\forall x.\varphi.$$

Being able to prove the Barcan biterm depends crucially on the constant domain assumption in the semantics, which leads to the particular form of the existential and universal rules in Figure B.3. For the varying domain semantics, the corresponding rules are more complicated and the Barcan biterm fails to hold generally.

Example 5.4.2. Suppose a theory consists of just the local assumption $\Box_i\forall x.\varphi$, where φ is some biterm. Then the proof in Figure 5.9 shows that $\forall x.\Box_i\varphi$ is a consequence of the theory.

An explanation of this proof is as follows. Item 1 is the negation of the biterm to be proved; 2 is from 1 by an existential rule; 3 is from 2 by a possibility rule; 4 is a local assumption; 5 is from 4 by a necessity rule; 6 is from 5 by a universal rule; now the branch closes by 3 and 6.

In effect, this proof establishes the converse Barcan biterm

$$\Box_i\forall x.\varphi \longrightarrow \forall x.\Box_i\varphi.$$

Example 5.4.3. Let α be a type, $p : \alpha \to \mathit{Bool}$ a predicate, and t a term of type α. Suppose a theory consists of just the local assumption $\Box_j(p\ t)$. Then the proof in Figure 5.10 shows that $\Box_j\forall x.(((=\ t)\ x) \longrightarrow (p\ x))$ is a consequence of the theory.

An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is from 1 by a possibility rule; 3 is from 2 by an existential rule; 4 and 5 are from 3 by a conjunctive rule; 6 is from 4 and 5 by the substitutivity rule; 7 is a local assumption; 8 is from 7 by a necessity rule; now the branch closes by 6 and 8.

$\underline{\blacklozenge((mean\ prices) < (mean \bullet prices))}$ [5.3.10]

$((mean\ \underline{prices}) < (mean \bullet prices)) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices))$ [5.3.1]

$(\underline{(mean\ (gaussian\ 400\ 20))} < (mean \bullet prices)) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices))$ [5.3.7]

$(400 < (mean \bullet \underline{prices})) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices))$ [5.3.2]

$(400 < (mean\ \underline{\bullet(gaussian\ 360\ 25)})) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices))$ [M]

$(400 < \underline{(mean\ (gaussian\ 360\ 25))}) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices))$ [5.3.7]

$\underline{(400 < 360)} \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices))$

$\underline{\bot \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices))}$ [O2]

$\bullet\underline{\blacklozenge((mean\ prices) < (mean \bullet prices))}$ [5.3.10]

$\bullet(((mean\ \underline{prices}) < (mean \bullet prices)) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [5.3.2]

$\bullet((\underline{(mean\ (gaussian\ 360\ 25))} < (mean \bullet prices)) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [5.3.7]

$\bullet((360 < (mean \bullet \underline{prices})) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [5.3.3]

$\bullet((360 < (mean\ \underline{\bullet\lambda x.if\ x = 300\ then\ 0.7\ else\ if\ x = 310\ then\ 0.3\ else\ 0})) \vee$
$\bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [M]

$\bullet((360 < \underline{(mean\ \lambda x.if\ x = 300\ then\ 0.7\ else\ if\ x = 310\ then\ 0.3\ else\ 0)}) \vee$
$\bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [5.3.9]

$\bullet((360 < \underline{300 \times 0.7} + (mean\ \lambda x.if\ x = 310\ then\ 0.3\ else\ 0)) \vee$
$\bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$

$\bullet((360 < 210 + \underline{(mean\ \lambda x.if\ x = 310\ then\ 0.3\ else\ 0)}) \vee$
$\bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [5.3.9]

$\bullet((360 < 210 + \underline{310 \times 0.3} + (mean\ \lambda x.0)) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$

$\bullet((360 < 210 + 93 + \underline{(mean\ \lambda x.0)}) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [5.3.8]

$\bullet((360 < \underline{210 + 93 + 0}) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$

$\bullet(\underline{(360 < 303)} \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$

$\bullet\underline{(\bot \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))}$ [O2]

$\bullet\bullet\underline{\blacklozenge((mean\ prices) < (mean \bullet prices))}$ [5.3.10]

$\bullet\bullet(((mean\ \underline{prices}) < (mean \bullet prices)) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [5.3.3]

$\vdots$

$\bullet\bullet((303 < (mean \bullet \underline{prices})) \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$ [5.3.4]

$\vdots$

$\bullet\bullet(\underline{(303 < 330)} \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))$

$\bullet\bullet\underline{(\top \vee \bullet\blacklozenge((mean\ prices) < (mean \bullet prices)))}$ [O1]

$\underline{\bullet\bullet\top}$ [M]

$\underline{\bullet\top}$ [M]

$\top$

Figure 5.7: Computation of $\blacklozenge((mean\ prices) < (mean \bullet prices))$

1	$\neg\Box_i\forall x.\varphi$	1.
1.1_i	$\neg\forall x.\varphi$	2.
1.1_i	$\neg\varphi\{x/y\}$	3.
1	$\forall x.\Box_i\varphi$	4.
1	$\Box_i\varphi\{x/y\}$	5.
1.1_i	$\varphi\{x/y\}$	6.

Figure 5.8: Proof of rank 0 of $\Box_i\forall x.\varphi$

1	$\neg\forall x.\Box_i\varphi$	1.
1	$\neg\Box_i\varphi\{x/y\}$	2.
1.1_i	$\neg\varphi\{x/y\}$	3.
1	$\Box_i\forall x.\varphi$	4.
1.1_i	$\forall x.\varphi$	5.
1.1_i	$\varphi\{x/y\}$	6.

Figure 5.9: Proof of rank 0 of $\forall x.\Box_i\varphi$

Example 5.4.4. Consider a formula φ containing the free variables $x_1, \ldots, x_n$ $(n \geq 0)$. Suppose that a theory consists of the local assumptions

$$\Box_{l_1} \cdots \Box_{l_p} \forall(\varphi)$$
$$\forall(\Box_{l_1} \cdots \Box_{l_p}\varphi \longrightarrow \Box_{j_1} \cdots \Box_{j_r}\varphi).$$

The proof of rank 0 of $\Box_{j_1} \cdots \Box_{j_r}\forall(\varphi)$ in Figure 5.11 shows that $\Box_{j_1} \cdots \Box_{j_r}\forall(\varphi)$ is a consequence of the theory. In Figure 5.11, φ' denotes the formula $\varphi\{x_1/y_1, \ldots, x_n/y_n\}$, where $y_1, \ldots, y_n$ are variables new to the branch of the proof.

An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is a local assumption; 3 is a local assumption; 4 is from 1 by a possibility rule; 5 is from 4 by possibility rules; 6 is from 5 by existential rules; 7 is from 3 by a universal rule; 8 and 14 are from 7 by a disjunctive rule; 9 is from 8 by a possibility rule; 10 is from 9 by possibility rules; 11 is from 2 by a necessity rule; 12 is from 11 by necessity rules; 13 is from 12 by universal rules; 15 is from 14 by a necessity rule; 16 is from 15 by necessity rules; now one branch closes by 10 and 13 and the other branch closes by 6 and 16.

1	$\neg\Box_j\forall x.(((=\ t)\ x) \longrightarrow (p\ x))$	1.
1.1_j	$\neg\forall x.(((=\ t)\ x) \longrightarrow (p\ x))$	2.
1.1_j	$\neg(((=\ t)\ y) \longrightarrow (p\ y))$	3.
1.1_j	$((=\ t)\ y)$	4.
1.1_j	$\neg(p\ y)$	5.
1.1_j	$\neg(p\ t)$	6.
1	$\Box_j(p\ t)$	7.
1.1_j	$(p\ t)$	8.

Figure 5.10: Proof of rank 0 of $\Box_j\forall x.(((=\ t)\ x) \longrightarrow (p\ x))$

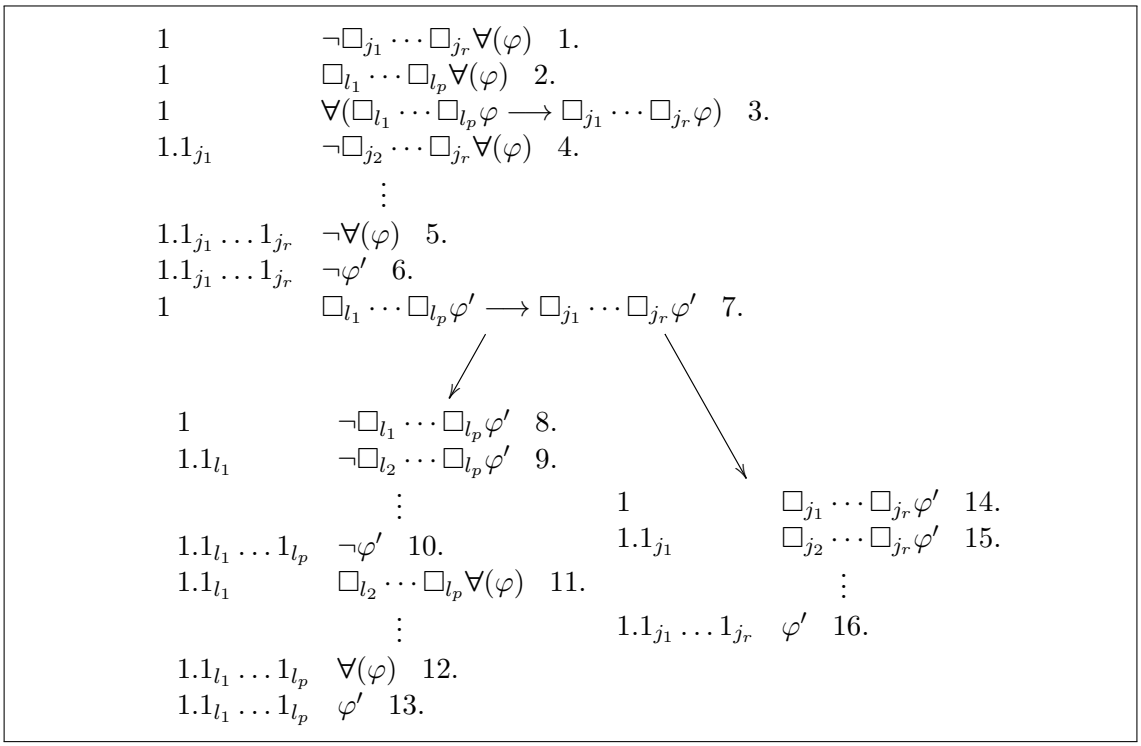

Figure 5.11: Proof of rank 0 of $\Box_{j_1}\cdots\Box_{j_r}\forall(\varphi)$

Example 5.4.5. This example comes from [8]. Consider the agents, Peter, John, and Wendy, with belief modalities $\boldsymbol{B}_p$, $\boldsymbol{B}_j$, and $\boldsymbol{B}_w$. Suppose that Peter believes that *time* is true, Peter believes that John believes that *place* is true, Wendy believes that if Peter believes that *time* is true, then John believes that *time* is true, and Peter believes that John believes that if *time* and *place* are true, then *appointment* is true. These belief formulas are captured in the following local assumptions of the belief theory.

$\boldsymbol{B}_p time$
$\boldsymbol{B}_p\boldsymbol{B}_j place$
$\boldsymbol{B}_w(\boldsymbol{B}_p time \longrightarrow \boldsymbol{B}_j time)$
$\boldsymbol{B}_p\boldsymbol{B}_j(place \wedge time \longrightarrow appointment)$.

Suppose, in addition, that if Peter believes something, then Peter believes that he believes that thing; everything believed by Wendy is believed by Peter; and if Peter believes that John believes something, then John believes that Peter believes the same thing. The local assumptions that capture all this are as follows.

$\boldsymbol{B}_p\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_p\boldsymbol{B}_p\boldsymbol{\varphi}$
$\boldsymbol{B}_w\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_p\boldsymbol{\varphi}$
$\boldsymbol{B}_p\boldsymbol{B}_j\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_j\boldsymbol{B}_p\boldsymbol{\varphi}$,

where $\boldsymbol{\varphi}$ is a syntactical variable standing for an arbitrary formula. The proof below uses the derived rules that correspond to these local implicational assumptions.

Suppose now that one wants to show that John believes that Peter believes *appointment* is true, that is,

$$\boldsymbol{B}_j\boldsymbol{B}_p \mathit{appointment},$$

is a theorem of the belief theory. The tableau proof of this formula is given in Figure 5.12.

An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 to 5 are local assumptions; 6 is from 1 by a derived rule from $\boldsymbol{B}_p\boldsymbol{B}_j\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_j\boldsymbol{B}_p\boldsymbol{\varphi}$; 7 is from 6 by a possibility rule; 8 is from 7 by a possibility rule; 9 is from 5 by a necessity rule; 10 is from 9 by a necessity rule; 11 and 25 are from 10 by a disjunctive rule; 12 and 15 are from 11 by a disjunctive rule; 13 is from 3 by a necessity rule; 14 is from 13 by a necessity rule; this branch now closes by 12 and 14; 16 is from 4 by a derived rule from $\boldsymbol{B}_w\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_p\boldsymbol{\varphi}$; 17 is from 16 by a necessity rule; 18 and 23 are from 17 by a disjunctive rule; 19 is from 18 by a possibility rule; 20 is from 2 by a derived rule from $\boldsymbol{B}_p\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}_p\boldsymbol{B}_p\boldsymbol{\varphi}$; 21 is from 20 by a necessity rule; 22 is from 21 by a necessity rule; this branch now closes by 19 and 22; 24 is from 23 by a necessity rule; this branch now closes by 15 and 24; and the remaining branch closes by 8 and 25.

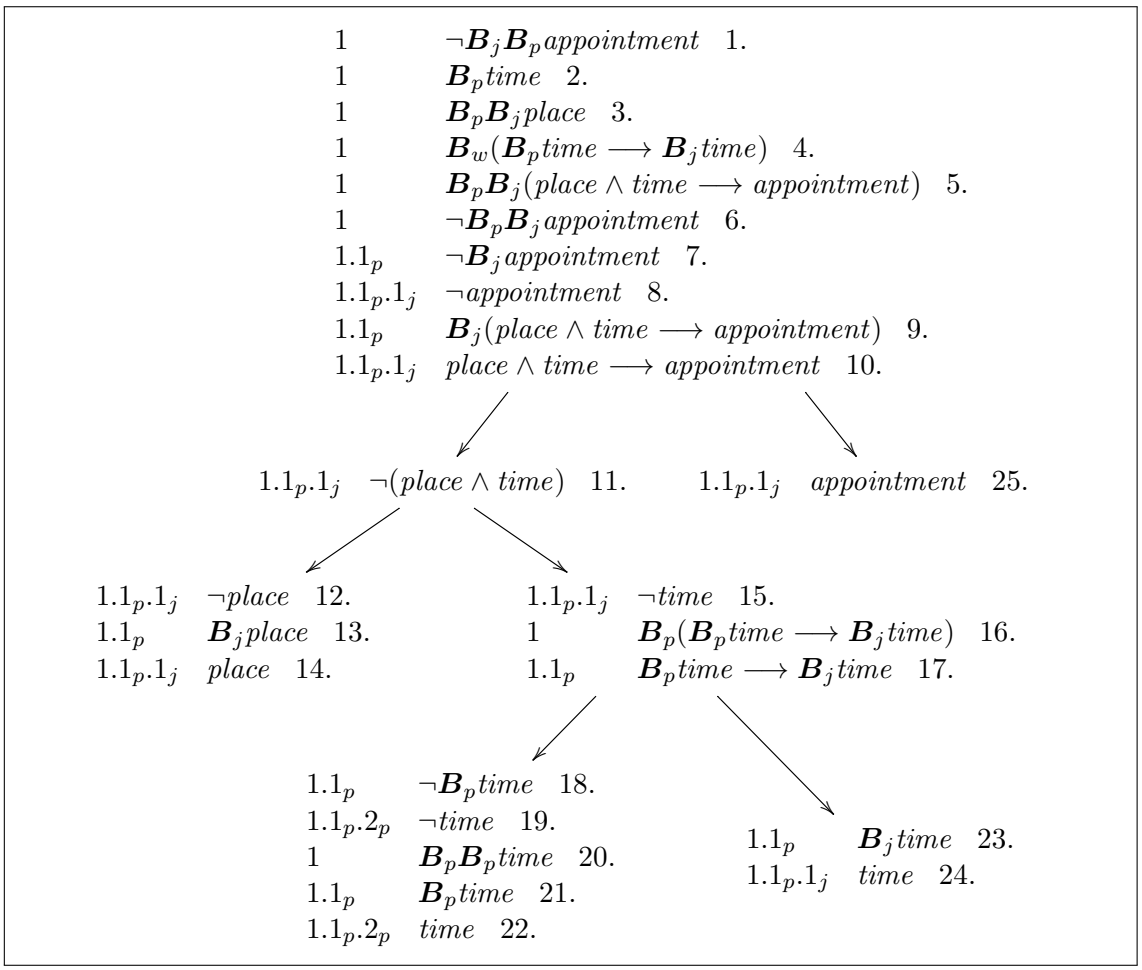

Figure 5.12: Proof of rank 0 of $\boldsymbol{B}_j\boldsymbol{B}_p \mathit{appointment}$

Example 5.4.6. Let $\mathfrak{T}$ be a theory, $q : \sigma \to Bool$, and $t \in \mathfrak{B}_\sigma$. It is shown that

$$\mathfrak{T} \vdash \Box_{j_1} \cdots \Box_{j_r} \forall x.(((= t)\ x) \longrightarrow (q\ x)) \text{ iff } \mathfrak{T} \vdash \Box_{j_1} \cdots \Box_{j_r} (q\ t).$$

This example is relevant to belief acquisition.

It suffices to show that

$$\vdash \Box_{j_1} \cdots \Box_{j_r} \forall x.(((= t)\ x) \longrightarrow (q\ x)) \longrightarrow \Box_{j_1} \cdots \Box_{j_r} (q\ t)$$

and

$$\vdash \Box_{j_1} \cdots \Box_{j_r} (q\ t) \longrightarrow \Box_{j_1} \cdots \Box_{j_r} \forall x.(((= t)\ x) \longrightarrow (q\ x)).$$

For the first of these, the proof is given in Figure 5.13. An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 and 3 are from 1 by a conjunctive rule; 4 is from 3 by a possibility rule; 5 is from 4 by possibility rules; 6 is from 2 by a necessity rule; 7 is from 6 by necessity rules; 8 is from 7 by a universal rule; 9 and 10 are from 8 by a disjunctive rule; 11 is the reflexivity rule; now the first branch closes by 9 and 11; and the second branch closes by 5 and 10.

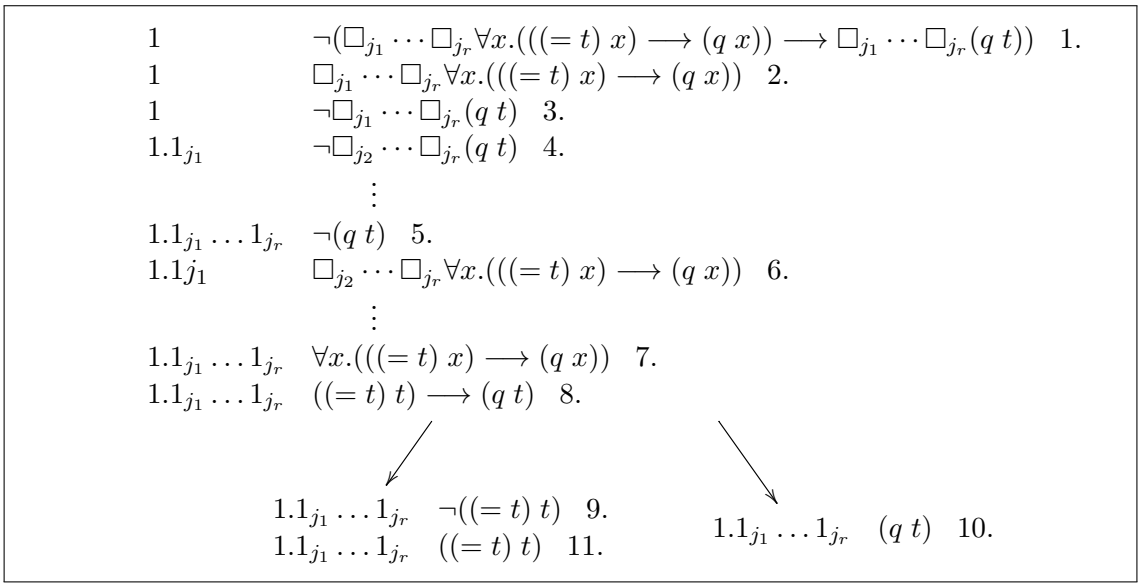

Figure 5.13: Proof of rank 0 of $\Box_{j_1} \cdots \Box_{j_r} \forall x.(((= t)\ x) \longrightarrow (q\ x)) \longrightarrow \Box_{j_1} \cdots \Box_{j_r} (q\ t)$

For the second, the proof is given in Figure 5.14. An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 and 3 are from 1 by a conjunctive rule; 4 is from 3 by a possibility rule; 5 is from 4 by possibility rules; 6 is from 5 by an existential rule; 7 and 8 are from 6 by a conjunctive rule; 9 is from 8 and 7 by the substitutivity rule; 10 is from 2 by a necessity rule; 11 is from 10 by necessity rules; now the branch closes by 9 and 11.

So far, the case where theories consist solely of formulas has been considered. Now the discussion turns to the more general case where theories can contain biterms that are not formulas and show why this generalisation is useful.

$$
\begin{array}{lll}
1 & \neg(\Box_{j_1}\cdots\Box_{j_r}(q\ t) \longrightarrow \Box_{j_1}\cdots\Box_{j_r}\forall x.(((=t)\ x) \longrightarrow (q\ x))) & 1.\\
1 & \Box_{j_1}\cdots\Box_{j_r}(q\ t) & 2.\\
1 & \neg\Box_{j_1}\cdots\Box_{j_r}\forall x.(((=t)\ x) \longrightarrow (q\ x)) & 3.\\
1.1_{j_1} & \neg\Box_{j_2}\cdots\Box_{j_r}\forall x.(((=t)\ x) \longrightarrow (q\ x)) & 4.\\
 & \vdots & \\
1.1_{j_1}\ldots 1_{j_r} & \neg\forall x.(((=t)\ x) \longrightarrow (q\ x)) & 5.\\
1.1_{j_1}\ldots 1_{j_r} & \neg(((=t)\ y) \longrightarrow (q\ y))) & 6.\\
1.1_{j_1}\ldots 1_{j_r} & ((=t)\ y) & 7.\\
1.1_{j_1}\ldots 1_{j_r} & \neg(q\ y) & 8.\\
1.1_{j_1}\ldots 1_{j_r} & \neg(q\ t) & 9.\\
1.1_{j_1} & \Box_{j_2}\cdots\Box_{j_r}(q\ t) & 10.\\
 & \vdots & \\
1.1_{j_1}\ldots 1_{j_r} & (q\ t) & 11.
\end{array}
$$

Figure 5.14: Proof of rank 0 of $\Box_{j_1}\cdots\Box_{j_r}(q\ t) \longrightarrow \Box_{j_1}\cdots\Box_{j_r}\forall x.(((=t)\ x) \longrightarrow (q\ x))$

Example 5.4.7. The discussion starts with a simple theory containing formulas. Consider the predicates $p, q, r : \sigma \to \mathit{Bool}$ and the theory having the following local assumptions.

$$\forall x.((p\ x) \longrightarrow (q\ x))$$
$$\forall x.((q\ x) \longrightarrow (r\ x)).$$

Figure 5.15 gives the proof of the theorem $\forall x.((p\ x) \longrightarrow (r\ x))$. The soundness result, Proposition B.3.5, shows that $\forall x.((p\ x) \longrightarrow (r\ x))$ is a consequence of theory.

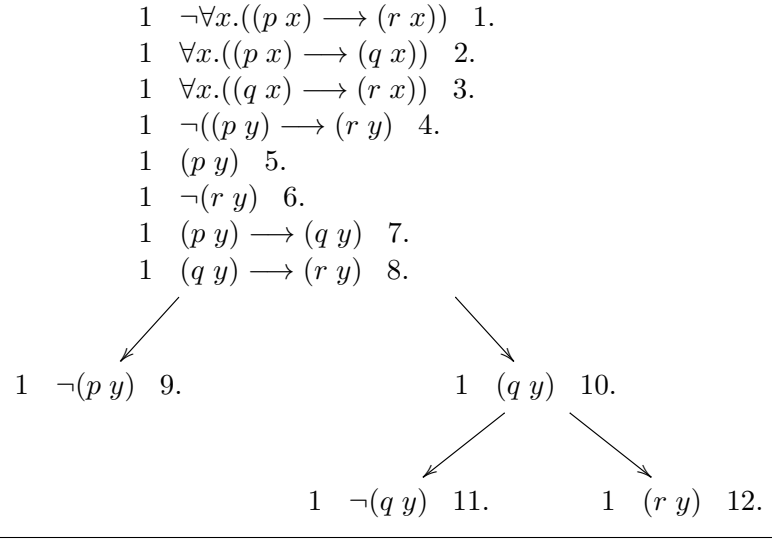

Figure 5.15: Proof of rank 0 of $\forall x.((p\ x) \longrightarrow (r\ x))$

The next step is to redo this example using biterms instead of formulas. Note that a biterm φ is valid at a world in an interpretation iff the formula $(\Pi\ \varphi)$ has the same

property, by Proposition B.2.2. Thus the implicit Π at the front of each formula in the preceding theory can be removed to obtain the following (essentially) equivalent theory.

$$\lambda x.((p\ x) \longrightarrow (q\ x))$$
$$\lambda x.((q\ x) \longrightarrow (r\ x)).$$

Similarly, the preceding theorem can be replaced by the biterm $\lambda x.((p\ x) \longrightarrow (r\ x))$. Figure 5.16 now gives the proof of this theorem from the biterm theory. The step that gives the biterms in items 4 and 5 in this proof uses the equation

$$\neg\lambda x.(\varphi \longrightarrow \psi) = \lambda x.\varphi \wedge \neg\lambda x.\psi,$$

which can be used as a local assumption, by Proposition B.2.1. Proposition B.3.5 shows that $\lambda x.((p\ x) \longrightarrow (r\ x))$ is a consequence of biterm theory, a statement equivalent to the analogous statement for the preceding version using formulas.

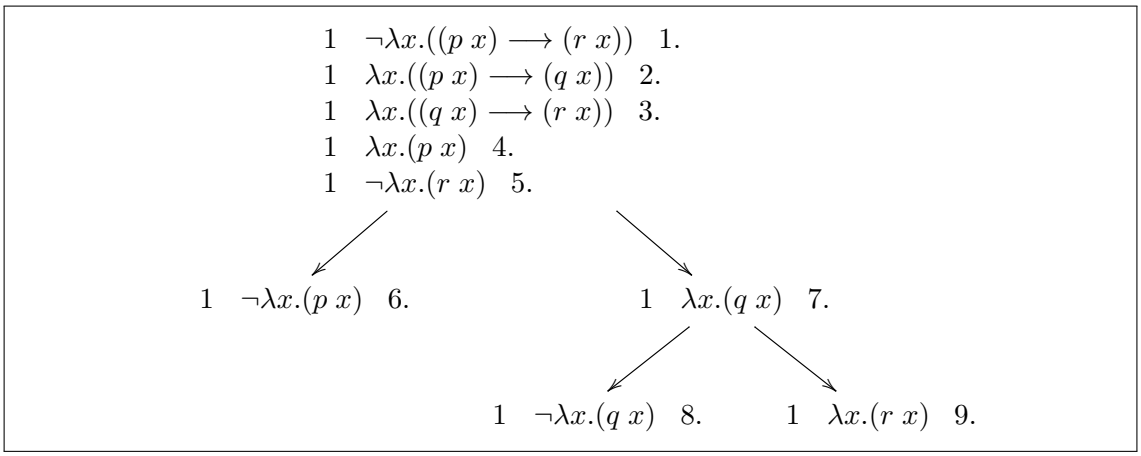

Figure 5.16: Proof of rank 0 of $\lambda x.((p\ x) \longrightarrow (r\ x))$

The biterm formulation is a little simpler than the formulation using formulas, but it can still be improved a lot. The point is that the use of variables complicates the theory and proof, and the variables are unnecessary. A better biterm formulation is the theory

$$p \longrightarrow q$$
$$q \longrightarrow r,$$

where $p \longrightarrow r$ is the theorem to be proved. Figure 5.17 contains this proof which is clearly as simple as any proof for this situation could possibly be. All reliance on variables has been eliminated and the proof is structurally equivalent to the proof for the case that p, q, and r are propositional constants (not predicate constants).

Biterms allow one to avoid the use of variables in knowledge representation and reasoning tasks when they are not really needed. Indeed, from the biterm perspective, many conventional uses of variables are simply gratuitous – they are only needed because, by the conventional definition, theories have to consist of formulas. Consequently, it is advocated that one should look for opportunities to use biterms that make knowledge representation

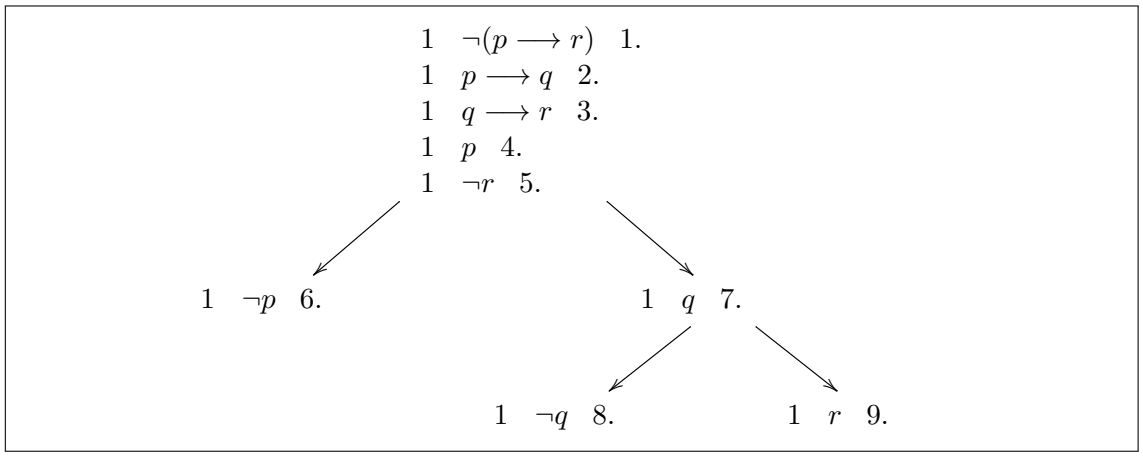

Figure 5.17: Proof of rank 0 of $p \longrightarrow r$

and reasoning tasks simpler. On the other hand, any situation that uses biterms can be forced into one that only uses formulas – simply apply Π to each assumption and to the biterm to be proved.

Here is an example of a propositional proof that extends immediately to the biterm case.

Example 5.4.8. Figure 5.18 gives a proof of the distribution axiom

$$\Box_i(\varphi \longrightarrow \psi) \longrightarrow (\Box_i\varphi \longrightarrow \Box_i\psi),$$

for the case that φ and ψ are biterms.

An explanation of this proof is as follows. Item 1 is the negation of the biterm to be proved; 2 and 3 are from 1 by a conjunctive rule; 4 and 5 are from 3 by a conjunctive rule; 6 is from 5 by a possibility rule; 7 is from 4 by a necessity rule; 8 is from 2 by a necessity rule; 9 and 10 are from 8 by a disjunctive rule; now the first branch closes by 7 and 9, and the second branch closes by 6 and 10.

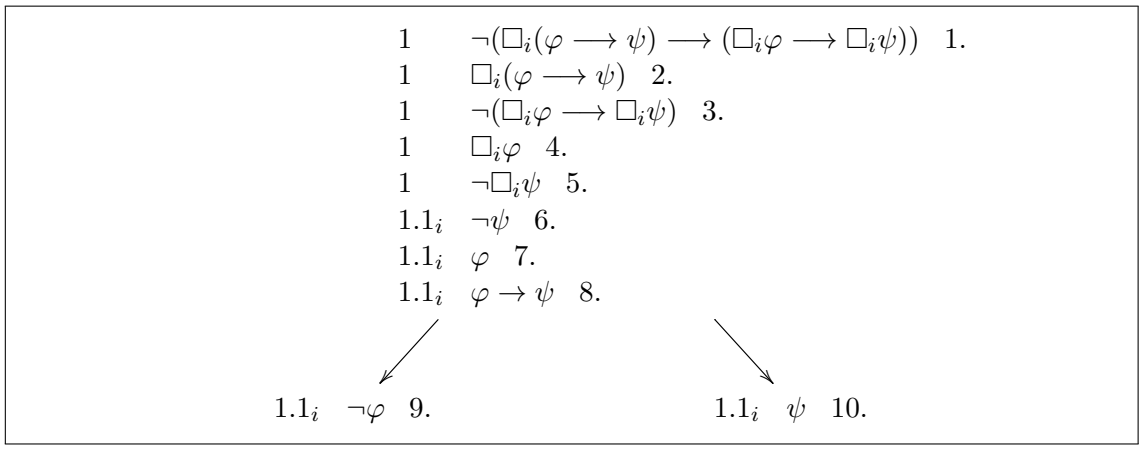

Figure 5.18: Proof of rank 0 of $\Box_i(\varphi \longrightarrow \psi) \longrightarrow (\Box_i\varphi \longrightarrow \Box_i\psi)$

It is possible to simulate a computation by a proof.

Example 5.4.9. Consider a theory that contains the local assumption

$$\boldsymbol{B}_i\boldsymbol{B}_j\forall x.((p\ x) = (q\ x) \wedge (r\ x))$$

and the one step computation

$$\lambda x.\boldsymbol{B}_j\underline{(p\ x)}$$
$$\lambda x.\boldsymbol{B}_j((q\ x) \wedge (r\ x))$$

using $\boldsymbol{B}_i$ which shows that the result

$$\boldsymbol{B}_i(\lambda x.\boldsymbol{B}_j(p\ x) = \lambda x.\boldsymbol{B}_j((q\ x) \wedge (r\ x)))$$

is a consequence of the theory.

Figure 5.19 is a proof that simulates this computation. An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is a local assumption; 3 is from 1 by a possibility rule; 4 is a global assumption (the axiom of extensionality); 5 is from 4 by universal rules; 6 is from 3 and 5 by the substitutivity rule; 7 is from 6 by an existential rule; 8 is a global assumption (β-reduction); 9 is a global assumption (β-reduction); 10 is from 7 and 8 by the substitutivity rule; 11 is from 9 and 10 by the substitutivity rule; 12 is the global assumption $\Box_i(\boldsymbol{s} = \boldsymbol{t}) \longrightarrow (\Box_i\boldsymbol{s} = \Box_i\boldsymbol{t})$; 13 and 17 are from 12 by a disjunctive rule; 14 is from 13 by a possibility rule; 15 is from 2 by necessity rules; 16 is from 15 by a universal rule; now the first branch closes by 14 and 16, and the second closes by 11 and 17.

This example illustrates just how much more complicated the proof simulation of a computation is compared with the computation itself.

5.5 Computation and Proof Examples

Here are several example to illustrate the combination of computation and proof, which is described in detail in Section B.3.3.

Example 5.5.1. Consider an alphabet containing the constants

$$\mathit{setExists}_1 : (\alpha \to \mathit{Bool}) \to \{\alpha\} \to \mathit{Bool}$$
$$\wedge_2 : (\alpha \to \mathit{Bool}) \to (\alpha \to \mathit{Bool}) \to \alpha \to \mathit{Bool}$$
$$\mathit{top} : \alpha \to \mathit{Bool}$$
$$\wedge : \mathit{Bool} \to \mathit{Bool} \to \mathit{Bool},$$

for some type α, and a theory $\mathcal{T}$ that includes

$$(\mathit{setExists}_1\ p\ t) = \exists x.((p\ x) \wedge (x \in t))$$
$$(\wedge_2\ p_1\ p_2\ x) = (p_1\ x) \wedge (p_2\ x)$$
$$(\mathit{top}\ x) = \top$$
$$x \wedge \top = x$$

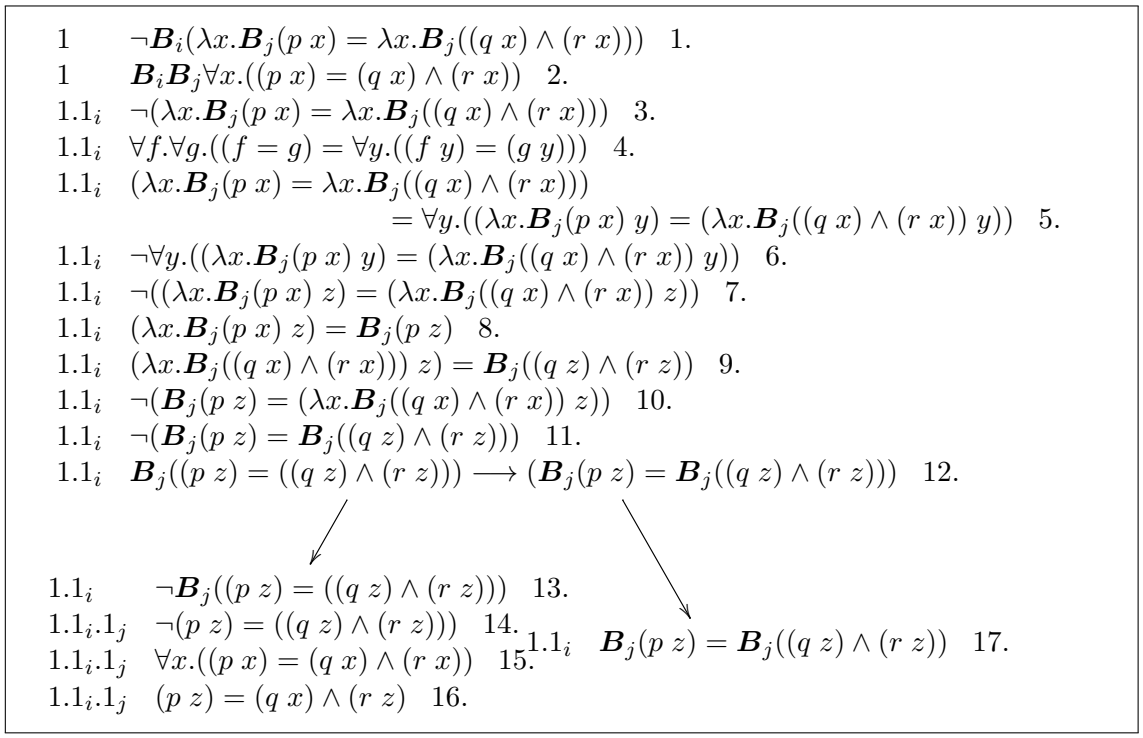

Figure 5.19: Proof of $\boldsymbol{B}_i(\lambda x.\boldsymbol{B}_j(p\ x) = \lambda x.\boldsymbol{B}_j((q\ x) \wedge (r\ x)))$

as global assumptions.

Let p and q be specific predicates on items of type α, and consider the two predicates

$$(\mathit{setExists}_1\ (\wedge_2\ p\ q))$$

and

$$(\mathit{setExists}_1\ (\wedge_2\ p\ \mathit{top})).$$

It will be shown that

$$\Box_j\forall x.((\mathit{setExists}_1\ (\wedge_2\ p\ q)\ x) \longrightarrow (\mathit{setExists}_1\ (\wedge_2\ p\ \mathit{top})\ x))$$

is a theorem of a proof of rank 1 with respect to $\mathcal{T}$ and, therefore, is a consequence of $\mathcal{T}$. The proof of this formula is given in Figure 5.20. In this proof, there are two subsidiary computations of rank 0: the computation of $(\mathit{setExists}_1\ (\wedge_2\ p\ q)\ y)$ is given in Figure 5.21 and the computation of $(\mathit{setExists}_1\ (\wedge_2\ p\ \mathit{top})\ y)$ is given in Figure 5.22. Each computation uses the leftmost selection rule and has the redexes underlined.

An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is from 1 by a possibility rule; 3 is from 2 by an existential rule; 4 and 5 are from 3 by a conjunctive rule; 6 is by introduction of the result from Figure 5.21; 7 is from 6 by a necessity rule; 8 is from 4 and 7 by the substitutivity rule; 9 is from 8 by an existential rule; 10, 11, and 12 are from 9 by conjunctive rules; 13 is by introduction of the result from Figure 5.22; 14 is from 13 by a necessity rule; 15 is from 5 and 14 by the substitutivity

$1 \quad \neg\Box_j \forall x.((setExists_1\ (\wedge_2\ p\ q)\ x) \longrightarrow (setExists_1\ (\wedge_2\ p\ top)\ x))$ 1.
$1.1_j \quad \neg\forall x.((setExists_1\ (\wedge_2\ p\ q)\ x) \longrightarrow (setExists_1\ (\wedge_2\ p\ top)\ x))$ 2.
$1.1_j \quad \neg((setExists_1\ (\wedge_2\ p\ q)\ y) \longrightarrow (setExists_1\ (\wedge_2\ p\ top)\ y))$ 3.
$1.1_j \quad setExists_1\ (\wedge_2\ p\ q)\ y$ 4.
$1.1_j \quad \neg(setExists_1\ (\wedge_2\ p\ top)\ y)$ 5.
$1 \quad \Box_j((setExists_1\ (\wedge_2\ p\ q)\ y) = \exists z.(((p\ z) \wedge (q\ z)) \wedge (z \in y)))$ 6.
$1.1_j \quad (setExists_1\ (\wedge_2\ p\ q)\ y) = \exists z.(((p\ z) \wedge (q\ z)) \wedge (z \in y))$ 7.
$1.1_j \quad \exists z.(((p\ z) \wedge (q\ z)) \wedge (z \in y))$ 8.
$1.1_j \quad ((p\ v) \wedge (q\ v)) \wedge (v \in y)$ 9.
$1.1_j \quad (p\ v)$ 10.
$1.1_j \quad (q\ v)$ 11.
$1.1_j \quad (v \in y)$ 12.
$1 \quad \Box_j((setExists_1\ (\wedge_2\ p\ top)\ y) = \exists z.((p\ z) \wedge (z \in y)))$ 13.
$1.1_j \quad (setExists_1\ (\wedge_2\ p\ top)\ y) = \exists z.((p\ z) \wedge (z \in y))$ 14.
$1.1_j \quad \neg\exists z.((p\ z) \wedge (z \in y))$ 15.
$1.1_j \quad \neg((p\ v) \wedge (v \in y))$ 16.

$1.1_j \quad \neg(p\ v)$ 17. $\qquad$ $1.1_j \quad \neg(v \in y)$ 18.

Figure 5.20: Proof of $\Box_j \forall x.((setExists_1\ (\wedge_2\ p\ q)\ x) \longrightarrow (setExists_1\ (\wedge_2\ p\ top)\ x))$

$\underline{(setExists_1\ (\wedge_2\ p\ q)\ y)}$
$\exists z.(\underline{((\wedge_2\ p\ q)\ z)} \wedge (z \in y))$
$\exists z.(((p\ z) \wedge (q\ z)) \wedge (z \in y))$

Figure 5.21: Computation using $\Box_j$ of $(setExists_1\ (\wedge_2\ p\ q)\ y)$

$\underline{(setExists_1\ (\wedge_2\ p\ top)\ y)}$
$\exists z.(\underline{((\wedge_2\ p\ top)\ z)} \wedge (z \in y))$
$\exists z.(((p\ z) \wedge \underline{(top\ z)}) \wedge (z \in y))$
$\exists z.(\underline{((p\ z) \wedge \top)} \wedge (z \in y))$
$\exists z.((p\ z) \wedge (z \in y))$

Figure 5.22: Computation using $\Box_j$ of $(setExists_1\ (\wedge_2\ p\ top)\ y)$

rule; 16 is from 15 by a universal rule; 17 and 18 are from 16 by a disjunctive rule; now one branch closes by 10 and 17 and the other branch closes by 12 and 18. The proof is now complete.

In Proposition B.2.24, it was shown semantically that $f = \lambda x.t$ and $\forall x.((f\ x) = t)$ are equivalent. The next example shows this proof-theoretically.

Example 5.5.2. Figure 5.23 gives a proof of rank 1 of $(f = \lambda x.t) = \forall x.((f\ x) = t)$, where f is a constant and t is a term. This proof uses the global assumption

$$\forall y.\forall z.((y = z) = \forall x.((y\ x) = (z\ x))),$$

which is the axiom of extensionality.

1	$\neg((f = \lambda x.t) = \forall x.((f\ x) = t))$	1.
1	$\forall y.\forall z.((y = z) = \forall x.((y\ x) = (z\ x)))$	2.
1	$(f = \lambda x.t) = \forall x.((f\ x) = (\lambda x.t\ x))$	3.
1	$\forall x.((f\ x) = (\lambda x.t\ x)) = \forall x.((f\ x) = t)$	4.
1	$(f = \lambda x.t) = \forall x.((f\ x) = t)$	5.

Figure 5.23: Proof of rank 1 of $(f = \lambda x.t) = \forall x.((f\ x) = t)$

An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is a global assumption; 3 is from 2 by a universal rule; 4 is by introduction of the result of a (one-step) computation; 5 is from 3 and 4 by the substitutivity rule; now the branch closes by 1 and 5.

The last example for this section illustrates reasoning about biterms.

Example 5.5.3. It is common to have to show that one standard predicate is stronger than another, because a standard predicate needs to be simplified or because the relationship needs to be established by a belief acquisition algorithm, for example.

Consider the standard predicates

$$\Box_2 f \diamond \Box_1 p$$
$$\Box_2 f \diamond \Box_1 q,$$

where $p : \sigma \to \mathit{Bool}$ and $q : \sigma \to \mathit{Bool}$ are defined by

$$\Box_1(p = r \wedge s \wedge t)$$
$$\Box_1(q = s \wedge t).$$

Figure 5.24 contains a proof that the first predicate is stronger than the second, that is, that

$$\Box_2 f \diamond \Box_1 p \longrightarrow \Box_2 f \diamond \Box_1 q$$

is a theorem. An explanation of this proof is as follows. Item 1 is the negation of the biterm to be proved; 2 and 3 are from 1 by a conjunctive rule; 4 is the result from the

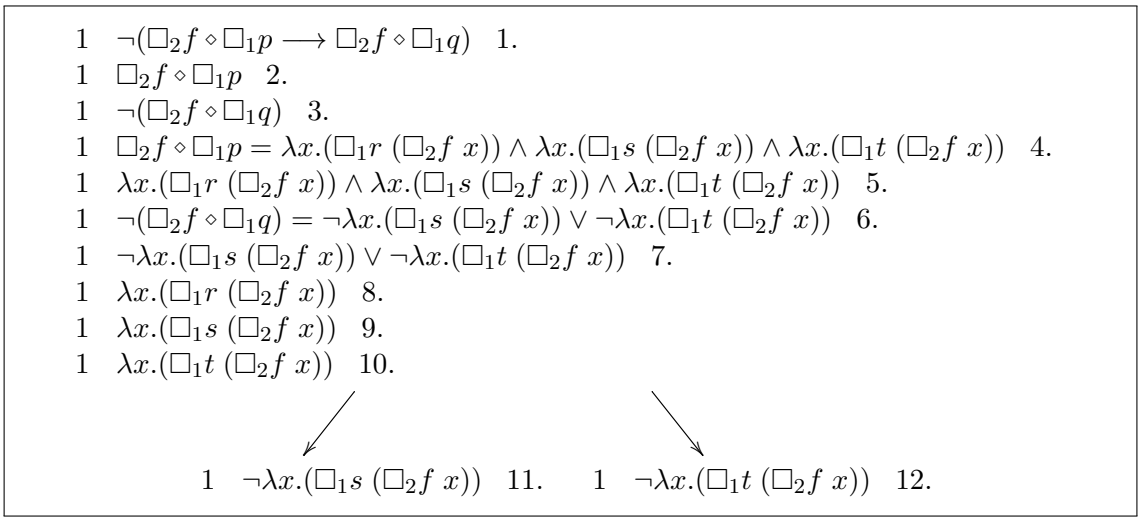

Figure 5.24: Proof of rank 1 of $\Box_2 f \diamond \Box_1 p \longrightarrow \Box_2 f \diamond \Box_1 q$

computation in Figure 5.25; 5 is from 2 and 4 by the substitutivity rule; 6 is the result from the computation in Figure 5.26; 7 is from 3 and 6 by the substitutivity rule; 8, 9, and 10 are from 5 by conjunctive rules; 11 and 12 are from 7 by a disjunctive rule; now the first branch closes by 9 and 11, and the second branch closes by 10 and 12.

The computations in Figures 5.25 and 5.26 use (amongst others) equations having the following form.

$$f \diamond g = \lambda x.(g\ (f\ x))$$
$$\Box_i(\varphi \wedge \psi) = \Box_i\varphi \wedge \Box_i\psi.$$

Actually, these two computations are not complete because the definitions for f, r, s and t have not been used. Clearly whether these definitions are used or not makes no essential difference to the proof, so the abbreviated computations correctly illustrate the ideas.

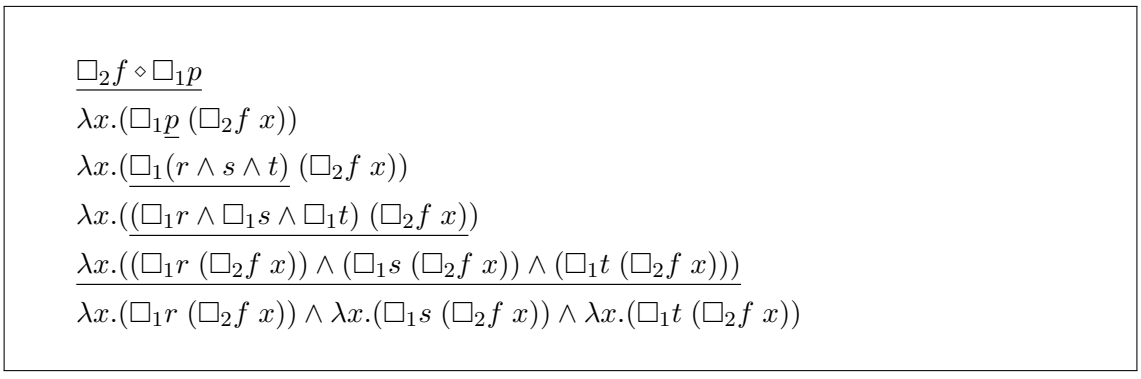

Figure 5.25: Computation of $\Box_2 f \diamond \Box_1 p$

Example 5.5.3 shows the general approach to showing that one standard predicate is stronger than another: simplify each predicate using a computation and then give the results of the computations to the theorem prover.

$$\neg(\underline{\Box_2 f \diamond \Box_1 q})$$
$$\neg\lambda x.(\Box_1 \underline{q}\ (\Box_2 f\ x))$$
$$\neg\lambda x.(\underline{\Box_1(s \wedge t)}\ (\Box_2 f\ x))$$
$$\neg\lambda x.(\underline{(\Box_1 s \wedge \Box_1 t)\ (\Box_2 f\ x)})$$
$$\neg\underline{\lambda x.((\Box_1 s\ (\Box_2 f\ x)) \wedge (\Box_1 t\ (\Box_2 f\ x)))}$$
$$\underline{\neg(\lambda x.(\Box_1 s\ (\Box_2 f\ x)) \wedge \lambda x.(\Box_1 t\ (\Box_2 f\ x)))}$$
$$\neg\lambda x.(\Box_1 s\ (\Box_2 f\ x)) \vee \neg\lambda x.(\Box_1 t\ (\Box_2 f\ x))$$

Figure 5.26: Computation of $\neg(\Box_2 f \diamond \Box_1 q)$

5.6 Reasoning about Beliefs

This section contains numerous examples of reasoning about beliefs, concentrating on non-empirical beliefs. Details about reasoning with empirical beliefs is postponed to the next section.

A convenient and common form for a belief is that of a piecewise-constant function. Consequently, the corresponding belief formula has a structure of the form

$$\begin{aligned}
&\square\forall x.((f\ x) = \\
&\qquad \mathit{if}\ (p_1\ x)\ \mathit{then}\ v_1 \\
&\qquad \mathit{else\ if}\ (p_2\ x)\ \mathit{then}\ v_2 \\
&\qquad\qquad \vdots \\
&\qquad \mathit{else\ if}\ (p_n\ x)\ \mathit{then}\ v_n \\
&\qquad \mathit{else}\ v_0),
\end{aligned} \tag{5.6.1}$$

where $\square$ is a (possibly empty) sequence of modalities, $f : \sigma \to \tau$, $p_1, \ldots, p_n$ are terms having type $\sigma \to \mathit{Bool}$, and $v_0, v_1, \ldots, v_n$ are terms of type τ. Here is an example that shows the usefulness of such belief formulas.

Example 5.6.1. This example concerns an infotainment agent, which is a multi-agent system that contains a number of agents with functionalities for recommending movies, TV programs, music and the like, as well as information agents with functionalities for searching for information on the Internet. Here, the discussion is centred on the TV recommender as a typical such agent.

First the types that will be needed and the data constructors corresponding to these types are introduced. Several standard types will be needed: *Bool* (the type of the booleans), *Nat* (the type of natural numbers), *Int* (the type of integers), and *String* (the type of strings). Also *List* denotes the (unary) list type constructor.

The following type synonyms are introduced.

$$State = Occurrence \times Status$$
$$Occurrence = Date \times Time \times Channel$$
$$Date = Day \times Month \times Year$$
$$Time = Hour \times Minute$$
$$Program = Title \times Subtitle \times Duration \times (List\ Genre) \times Classification \times Synopsis$$
$$Text = List\ String.$$

In addition, *Title*, *Subtitle*, and *Synopsis* are all defined to be *String*, and *Year*, *Month*, *Day*, *Hour*, *Minute* and *Duration* are all defined to be *Nat*.

The data constructors for the type *Status* are as follows.

$$Unknown, Yes, No : Status.$$

The meaning of *Unknown* is that a recommendation (about a program having a particular occurrence) hasn't yet been made, *Yes* means that it has a positive recommendation, and *No* means that it has a negative recommendation.

In the following, basic terms are used to represent individuals. So, for example, basic terms of type *Program* are used to represent TV programs and basic terms of type *Occurrence* are used to represent occurrences (of programs).

The constant of primary interest in this example is

$$uaccept : State \rightarrow Bool,$$

whose value is true if and only if the user is willing to watch television during the time when the program having the occurrence component of the state is on. The constant *uaccept* is a predicate that provides a state feature that is used by the TV agent with two other state features to determine whether or not to recommend a TV program.

Suppose that the theory $\mathcal{T}$, the belief theory of the TV agent, includes the following definitions as global assumptions.

$$add : Time \times Duration \rightarrow Time$$
$$\forall h.\ \forall m.\ \forall d.$$
$$((add\ ((h, m), d)) =$$
$$((60 \times h + m + d)\ div\ 60, (60 \times h + m + d)\ mod\ 60)).$$

$$proj_{Occurrence} : State \rightarrow Occurrence$$
$$\forall o.\ \forall s.$$
$$((proj_{Occurrence}\ (o, s)) = o).$$

$$proj_{Duration} : Program \rightarrow Duration$$
$$\forall t.\ \forall t'.\forall d.\forall g.\forall c.\forall s.$$
$$((proj_{Duration}\ (t, t', d, g, c, s)) = d).$$

Other global assumptions are definitions for the constants *weekday* (a predicate to determine whether a day is a weekday or not), $proj_{Hour}$ (the projection onto the *Hour* component of *Time*), $proj_{Time}$ (the projection onto the *Time* component of *Occurrence*), and *if_then_else*. Each of the global assumptions is true at each world of the intended pointed interpretation.

Suppose that $\mathcal{T}$ includes the following definitions as local assumptions. For these, $\boldsymbol{B}_u$ is the belief modality for the user and $\boldsymbol{B}_t$ is the belief modality for the TV agent.

$$
\begin{array}{l}
uaccept : State \rightarrow Bool \\
\boldsymbol{B}_t\ \forall x. \\
\quad ((uaccept\ x) = \\
\quad\quad \boldsymbol{B}_u(tv\_time\_acceptable\ (period\ (proj_{Occurrence}\ x)))).
\end{array}
$$

$$
\begin{array}{l}
period : Occurrence \rightarrow Date \times Time \times Time \\
\boldsymbol{B}_t\ \forall x. \\
\quad ((period\ x) = \\
\quad\quad ((proj_{Date}\ x), (proj_{Time}\ x), (add\ ((proj_{Time}\ x), (proj_{Duration}\ (tv\_guide\ x)))))).
\end{array}
$$

$$
\begin{array}{l}
tv\_guide : Occurrence \rightarrow Program \\
\boldsymbol{B}_t\ \forall x. \\
\quad ((tv\_guide\ x) = \\
\quad\quad if\ ((= ((21, 7, 2004), (19, 30), Win))\ x) \\
\quad\quad\quad\quad\quad\quad then\ (\text{“}Seinfeld\text{”}, \text{“”}, 30, [Sitcom], PG, \text{“}Kramer \ldots\text{”}) \\
\quad\quad else\ if\ ((= ((20, 7, 2004), (20, 30), ABC))\ x) \\
\quad\quad\quad\quad\quad\quad then\ (\text{“}The\ Bill\text{”}, \text{“”}, 50, [Drama], M, \text{“}Sun\ Hill \ldots\text{”}) \\
\quad\quad\quad\quad \vdots \\
\quad\quad else\ (\text{“ ”}, \text{“ ”}, 0, [], NA, \text{“ ”})).
\end{array}
$$

$$
\begin{array}{l}
tv\_time\_acceptable : Date \times Time \times Time \rightarrow Bool \\
\boldsymbol{B}_t\boldsymbol{B}_u\ \forall x. \\
\quad ((tv\_time\_acceptable\ x) = \\
\quad\quad if\ (weekday\ (proj_{Day}\ x)) \wedge (((proj_{Hour}\ (proj_{Start}\ x)) \geq 20) \wedge \\
\quad\quad\quad\quad\quad\quad\quad\quad ((proj_{Hour}\ (proj_{End}\ x)) \leq 23))\ then\ \top \\
\quad\quad else\ if\ \neg(weekday\ (proj_{Day}\ x)) \wedge (((proj_{Hour}\ (proj_{Start}\ x)) \geq 12) \wedge \\
\quad\quad\quad\quad\quad\quad\quad\quad ((proj_{Hour}\ (proj_{End}\ x)) \leq 25))\ then\ \top \\
\quad\quad else\ \bot).
\end{array}
$$

The term (“ ”, “ ”, 0, [], *NA*, “ ”) is the default value of type *Program*. The constant *tv_time_acceptable* is acquired by a belief acquisition process using training examples provided by the user. Hence the modal sequence $\boldsymbol{B}_t\boldsymbol{B}_u$ at the front of the definition: the

TV agent believes that the user believes that certain times are acceptable for watching television. The constant *tv_guide* is similarly obtained, but from an agent (the provider of television guides) that is not modelled here. Hence there is just the modality $\boldsymbol{B}_t$ at the front of the definition.

Suppose also that there is a local assumption

$$\boldsymbol{B}_t \forall x.((tv\_guide\ x) = \mathbf{u}) \longrightarrow \boldsymbol{B}_t \boldsymbol{B}_u \forall x.((tv\_guide\ x) = \mathbf{u}),$$

where $\mathbf{u}$ is a syntactical variable ranging over terms of type *Program*. The intuitive meaning of this scheme is that "if the TV agent believes the TV guide, then the TV agent believes the user believes the TV guide". This assumption is justified because it is natural for the TV agent to assume that the user has the same belief formulas about the TV guide as the TV agent does; after all, the TV guide is publicly available on the Web and in newspapers, and all versions of it are (nearly) identical.

Now let φ be the scope of the modality in the definition of *tv_guide*, that is, φ is the formula

$$\begin{array}{l}
\forall x. \\
\quad ((tv_guide\ x) = \\
\quad\quad \mathit{if}\ ((= ((21,7,2004),(19,30),\mathit{Win}))\ x) \\
\quad\quad\quad\quad\quad\quad \mathit{then}\ (\text{“}\mathit{Seinfeld}\text{”}, \text{“”}, 30, [\mathit{Sitcom}], \mathit{PG}, \text{“}\mathit{Kramer}\ldots\text{”}) \\
\quad\quad \mathit{else\ if}\ ((= ((20,7,2004),(20,30),\mathit{ABC}))\ x) \\
\quad\quad\quad\quad\quad\quad \mathit{then}\ (\text{“}\mathit{The\ Bill}\text{”}, \text{“”}, 50, [\mathit{Drama}], M, \text{“}\mathit{Sun\ Hill}\ldots\text{”}) \\
\quad\quad\quad \vdots \\
\quad\quad \mathit{else}\ (\text{“ ”}, 0, \text{“ ”}, \text{“ ”}, [], \mathit{NA}, \text{“ ”})).
\end{array}$$

Then, using as local assumptions the definition of *tv_guide* (that is, $\boldsymbol{B}_t\varphi$) and the instance $\boldsymbol{B}_t\varphi \longrightarrow \boldsymbol{B}_t\boldsymbol{B}_u\varphi$ of the above local assumption, Figure 5.27 gives a proof of rank 0 of the theorem $\boldsymbol{B}_t\boldsymbol{B}_u\varphi$, that is,

$$\begin{array}{l}
\boldsymbol{B}_t\boldsymbol{B}_u\ \forall x. \\
\quad ((tv_guide\ x) = \\
\quad\quad \mathit{if}\ ((= ((21,7,2004),(19,30),\mathit{Win}))\ x) \\
\quad\quad\quad\quad\quad\quad \mathit{then}\ (\text{“}\mathit{Seinfeld}\text{”}, \text{“”}, 30, [\mathit{Sitcom}], \mathit{PG}, \text{“}\mathit{Kramer}\ldots\text{”}) \\
\quad\quad \mathit{else\ if}\ ((= ((20,7,2004),(20,30),\mathit{ABC}))\ x) \\
\quad\quad\quad\quad\quad\quad \mathit{then}\ (\text{“}\mathit{The\ Bill}\text{”}, \text{“”}, 50, [\mathit{Drama}], M, \text{“}\mathit{Sun\ Hill}\ldots\text{”}) \\
\quad\quad\quad \vdots \\
\quad\quad \mathit{else}\ (\text{“ ”}, \text{“ ”}, 0, [], \mathit{NA}, \text{“ ”})).
\end{array}$$

An explanation of the proof in Figure 5.27 is as follows. Item 1 is the negation of the formula to be proved; 2 is a local assumption; 3 is from 1 by a derived rule from $\boldsymbol{B}_t\varphi \longrightarrow \boldsymbol{B}_t\boldsymbol{B}_u\varphi$; now the branch closes by 2 and 3.

Of course, as an alternative to this approach to handling the TV guide, it would be possible to explicitly maintain both $\boldsymbol{B}_t\varphi$ and $\boldsymbol{B}_t\boldsymbol{B}_u\varphi$ in the belief theory of the TV agent.

1	$\neg \boldsymbol{B}_t \boldsymbol{B}_u \varphi$	1.
1	$\boldsymbol{B}_t \varphi$	2.
1	$\neg \boldsymbol{B}_t \varphi$	3.

Figure 5.27: Proof of rank 0 of $\boldsymbol{B}_t \boldsymbol{B}_u \varphi$

But this is not a good solution because the TV guide is regularly updated, so keeping both of these would lead to duplicated work and consistency problems. Thus it is better to maintain just $\boldsymbol{B}_t \varphi$ and infer $\boldsymbol{B}_t \boldsymbol{B}_u \varphi$ from this.

Suppose finally that there is a local assumption

$$\boldsymbol{B}_t \forall x.((\mathit{period}\ x) = \mathbf{u}) \longrightarrow \boldsymbol{B}_t \boldsymbol{B}_u \forall x.((\mathit{period}\ x) = \mathbf{u}),$$

where $\mathbf{u}$ is a syntactical variable ranging over terms of type $\mathit{Date} \times \mathit{Time} \times \mathit{Time}$. In a similar way, there is a proof of rank 0 of the theorem

$$\begin{aligned} &\boldsymbol{B}_t \boldsymbol{B}_u\ \forall x. \\ &\quad ((\mathit{period}\ x) = \\ &\qquad ((\mathit{proj}_{\mathit{Date}}\ x), (\mathit{proj}_{\mathit{Time}}\ x), (\mathit{add}\ ((\mathit{proj}_{\mathit{Time}}\ x), (\mathit{proj}_{\mathit{Duration}}\ (\mathit{tv\_guide}\ x)))))). \end{aligned}$$

Now, using these two theorems, Figure 5.28 gives a computation of rank 1 using $\boldsymbol{B}_t$ of $(\mathit{uaccept}\ (((20, 7, 2004), (20, 30), \mathit{ABC}), \mathit{Unknown}))$. It follows from Proposition B.3.10 that

$$\boldsymbol{B}_t((\mathit{uaccept}\ (((20, 7, 2004), (20, 30), \mathit{ABC}), \mathit{Unknown})) = \top)$$

is a consequence of $\mathcal{T}$. Informally, this states that the TV agent believes that the user is willing to watch television during the time when the program on channel ABC that starts at 8.30pm on 20th July, 2004, is on.

Example 5.6.2. This example illustrates the use of temporal modalities with predicates.

Suppose the belief theory of an agent is based on an alphabet that includes the following constants.

$$\begin{aligned} &p : \sigma \to \mathit{Bool} \\ &q : \sigma \to \mathit{Bool} \\ &A, B, \ldots, Z : \sigma, \end{aligned}$$

where $A, B, \ldots, Z$, are rigid. Suppose the belief theory includes the following local assumptions.

$$\begin{aligned} &\boldsymbol{B}\ \forall x.((p\ x) = \mathit{\ if\ } x = A \mathit{\ then\ } \top \mathit{\ else\ if\ } x = B \mathit{\ then\ } \top \mathit{\ else\ } \bot) \\ &\boldsymbol{B}\ \forall x.((q\ x) = \mathit{\ if\ } x = A \mathit{\ then\ } \top \mathit{\ else\ if\ } x = C \mathit{\ then\ } \top \mathit{\ else\ } \bot) \\ &\bullet \boldsymbol{B}\ \forall x.((p\ x) = \mathit{\ if\ } x = A \mathit{\ then\ } \top \mathit{\ else\ if\ } x = C \mathit{\ then\ } \top \mathit{\ else\ } \bot) \\ &\bullet \boldsymbol{B}\ \forall x.((q\ x) = \mathit{\ if\ } x = A \mathit{\ then\ } \top \mathit{\ else\ } \bot) \\ &\bullet^2 \boldsymbol{B}\ \forall x.((p\ x) = \mathit{\ if\ } x = C \mathit{\ then\ } \top \mathit{\ else\ } \bot) \\ &\bullet^2 \boldsymbol{B}\ \forall x.((q\ x) = \mathit{\ if\ } x = A \mathit{\ then\ } \top \mathit{\ else\ if\ } x = B \mathit{\ then\ } \top \mathit{\ else\ } \bot) \\ &\bullet^3 \boldsymbol{B}\ \forall x.((p\ x) = \bot) \\ &\bullet^3 \boldsymbol{B}\ \forall x.((q\ x) = \bot). \end{aligned}$$

$\underline{(uaccept\ (((20,7,2004),(20,30),ABC),Unknown))}$

$\boldsymbol{B}_u(tv\_time\_acceptable\ (period\ \underline{(proj_{Occurrence}\ (((20,7,2004),(20,30),ABC),Unknown))}))$

$\boldsymbol{B}_u(tv\_time\_acceptable\ \underline{(period\ ((20,7,2004),(20,30),ABC))})$

$\vdots$

$\boldsymbol{B}_u(tv\_time\_acceptable\ ((20,7,2004),(20,30),$
$(add\ ((20,30),(proj_{Duration}\ \underline{(tv\_guide\ ((20,7,2004),(20,30),ABC))})))))$

$\vdots$

$\boldsymbol{B}_u(tv\_time\_acceptable\ ((20,7,2004),(20,30),$
$(add\ ((20,30),\underline{(proj_{Duration}\ (\text{“}The\ Bill\text{”},\text{“”},50,[Drama],M,\text{“}Sun\ Hill\ldots\text{”}))})))))$

$\boldsymbol{B}_u\ (tv\_time\_acceptable\ ((20,7,2004),(20,30),\underline{(add\ ((20,30),50))}))$

$\vdots$

$\boldsymbol{B}_u\underline{(tv\_time\_acceptable\ ((20,7,2004),(20,30),(21,20)))}$

$\boldsymbol{B}_u(if\ \underline{(weekday\ ((20,7,2004)))} \wedge (((proj_{Hour}\ (20,30)) \geq 20) \wedge ((proj_{Hour}\ (21,20)) \leq 23))\ then\ \top$
$else\ if\ \neg(weekday\ (20,7,2004)) \wedge (((proj_{Hour}\ (20,30)) \geq 12) \wedge ((proj_{Hour}\ (21,20)) \leq 25))\ then\ \top$
$else\ \bot))$

$\vdots$

$\boldsymbol{B}_u(if\ \underline{\top \wedge (((proj_{Hour}\ (20,30)) \geq 20) \wedge ((proj_{Hour}\ (21,20)) \leq 23))}\ then\ \top$
$else\ if\ \neg(weekday\ (20,7,2004)) \wedge (((proj_{Hour}\ (20,30)) \geq 12) \wedge ((proj_{Hour}\ (21,20)) \leq 25))\ then\ \top$
$else\ \bot))$

$\vdots$

$\boldsymbol{B}_u\underline{(if\ \top\ then\ \top}$
$\underline{else\ if\ \neg(weekday\ (20,7,2004)) \wedge ((proj_{Hour}\ (20,30)) \geq 12) \wedge ((proj_{Hour}\ (21,20)) \leq 25)\ then\ \top}$
$\underline{else\ \bot))}$

$\underline{\boldsymbol{B}_u\top}$

$\top$

Figure 5.28: Computation of $(uaccept\ (((20,7,2004),(20,30),ABC),Unknown))$ of rank 1 using $\boldsymbol{B}_t$

The above is the kind of belief theory of an agent that remembers (some of) the past and uses a belief acquisition process to acquire new belief formulas. At time 0, the definitions of p and q make them false everywhere. At time 1, some training examples arrive and the new definitions of p and q that result from a belief acquisition process using these training examples are

$$\boldsymbol{B}\ \forall x.((p\ x) = \mathit{if}\ x = C\ \mathit{then}\ \top\ \mathit{else}\ \bot)$$
$$\boldsymbol{B}\ \forall x.((q\ x) = \mathit{if}\ x = A\ \mathit{then}\ \top\ \mathit{else\ if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \bot).$$

Meanwhile, at time 1, the old definitions for p and q have a $\bullet$ modality put in front of them to indicate that they are true at the last time. Thus they become as follows.

$$\bullet\boldsymbol{B}\ \forall x.((p\ x) = \bot)$$
$$\bullet\boldsymbol{B}\ \forall x.((q\ x) = \bot).$$

After two more time steps of belief acquisition, one arrives at the belief theory given at the beginning of this example.

Suppose also that the belief theory includes the following global assumptions.

$$\boldsymbol{\varphi}\,\boldsymbol{S}\,\boldsymbol{\psi} = \boldsymbol{\psi} \vee (\boldsymbol{\varphi} \wedge \bullet(\boldsymbol{\varphi}\,\boldsymbol{S}\,\boldsymbol{\psi}))$$
$$\bullet\boldsymbol{B}\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}\bullet\boldsymbol{\varphi}.$$

Now using the global assumption

$$\bullet\boldsymbol{B}\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}\bullet\boldsymbol{\varphi}$$

and the local assumption

$$\bullet\boldsymbol{B}\ \forall x.((q\ x) = \mathit{if}\ x = A\ \mathit{then}\ \top\ \mathit{else}\ \bot),$$

it is easy to show that

$$\boldsymbol{B}\bullet\ \forall x.((q\ x) = \mathit{if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \bot)$$

is a theorem of the belief theory.

Similarly, from the local assumption

$$\bullet^2\boldsymbol{B}\ \forall x.((q\ x) = \mathit{if}\ x = A\ \mathit{then}\ \top\ \mathit{else\ if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \bot),$$

Figure 5.29 shows that

$$\boldsymbol{B}\bullet^2\ \forall x.((q\ x) = \mathit{if}\ x = \ \mathit{then}\ \top\ \mathit{else\ if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \bot)$$

is a theorem of the belief theory. An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is a local assumption; 3 is from 1 by a derived rule from the global implicational assumption $\bullet\boldsymbol{B}\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}\bullet\boldsymbol{\varphi}$; 4 is from 3 by a possibility rule; 5 is from 4 by a derived rule from $\bullet\boldsymbol{B}\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}\bullet\boldsymbol{\varphi}$; 6 is from 2 by a necessity rule; the branch now closes by 5 and 6.

$1 \quad \neg \boldsymbol{B} \bullet^2 \varphi \quad 1.$
$1 \quad \bullet^2 \boldsymbol{B} \varphi \quad 2.$
$1 \quad \neg \bullet \boldsymbol{B} \bullet \varphi \quad 3.$
$1.1_\bullet \quad \neg \boldsymbol{B} \bullet \varphi \quad 4.$
$1.1_\bullet \quad \neg \bullet \boldsymbol{B} \varphi \quad 5.$
$1.1_\bullet \quad \bullet \boldsymbol{B} \varphi \quad 6.$

Figure 5.29: Proof of rank 0 of $\boldsymbol{B} \bullet^2 \varphi$, where φ is $\forall x.((q\ x) =$ *if* $x = A$ *then* $\top$ *else if* $x = B$ *then* $\top$ *else* $\bot)$

$$\underline{(p\ A)\ \boldsymbol{S}\ (q\ B)}$$
$$\underline{(q\ B)} \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B)))$$
$$\vdots$$
$$\underline{\bot \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B)))}$$
$$\underline{(p\ A)} \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B))$$
$$\vdots$$
$$\underline{\top \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B))}$$
$$\bullet\underline{((p\ A)\ \boldsymbol{S}\ (q\ B))}$$
$$\bullet(\underline{(q\ B)} \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B))))$$
$$\vdots$$
$$\bullet\underline{(\bot \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B))))}$$
$$\bullet(\underline{(p\ A)} \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B)))$$
$$\vdots$$
$$\bullet\underline{(\top \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B)))}$$
$$\bullet^2\underline{((p\ A)\ \boldsymbol{S}\ (q\ B))}$$
$$\bullet^2(\underline{(q\ B)} \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B))))$$
$$\vdots$$
$$\bullet^2\underline{(\top \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ B))))}$$
$$\underline{\bullet\bullet\top}$$
$$\underline{\bullet\top}$$
$$\top$$

Figure 5.30: Computation of rank 1 using $\boldsymbol{B}$ of $(p\ A)\ \boldsymbol{S}\ (q\ B)$

$\underline{(p\ A)\ \boldsymbol{S}\ (q\ D)}$

$\underline{(q\ D)} \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D)))$

$\vdots$

$\underline{\bot \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D)))}$

$\underline{(p\ A)} \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D))$

$\vdots$

$\underline{\top \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D))}$

$\bullet\underline{((p\ A)\ \boldsymbol{S}\ (q\ D))}$

$\bullet(\underline{(q\ D)} \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D))))$

$\vdots$

$\bullet\underline{(\bot \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D))))}$

$\bullet(\underline{(p\ A)} \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D)))$

$\vdots$

$\bullet\underline{(\top \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D)))}$

$\bullet^2\underline{((p\ A)\ \boldsymbol{S}\ (q\ D))}$

$\bullet^2(\underline{(q\ D)} \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D))))$

$\vdots$

$\bullet^2\underline{(\bot \vee ((p\ A) \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D))))}$

$\bullet^2(\underline{(p\ A)} \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D)))$

$\vdots$

$\bullet^2\underline{(\bot \wedge \bullet((p\ A)\ \boldsymbol{S}\ (q\ D)))}$

$\underline{\bullet\bullet\bot}$

$\underline{\bullet\bot}$

$\bot$

Figure 5.31: Computation of rank 1 using $\boldsymbol{B}$ of $(p\ A)\ \boldsymbol{S}\ (q\ D)$

Using theorems such as these, Figure 5.30 gives a computation of rank 1 using $\boldsymbol{B}$ of $(p\ A)\ \boldsymbol{S}\ (q\ B)$. This computation shows that the result $\boldsymbol{B}((p\ A)\ \boldsymbol{S}\ (q\ B) = \top)$ is a consequence of the belief theory.

It is easy to imagine the formula $\boldsymbol{B}((p\ A)\ \boldsymbol{S}\ (q\ B))$ being a useful condition in some application: if the agent believes that $(p\ A)$ has held ever since $(q\ B)$ held at some time in the past, then perform a certain action; else perform a different action.

By way of contrast, Figure 5.31 shows a computation that ends in $\bot$.

One potential problem in using a global assumption such as $\boldsymbol{\varphi}\,\boldsymbol{S}\,\boldsymbol{\psi} = \boldsymbol{\psi} \vee (\boldsymbol{\varphi} \wedge \bullet(\boldsymbol{\varphi}\,\boldsymbol{S}\,\boldsymbol{\psi}))$ in computations is that it might allow them to run forever. Thus one must be careful to ensure that the belief theory is such that any computation using this assumption will indeed terminate. For computations similar to Figures 5.30 and 5.31 that go into the past, one can ensure termination by having belief formulas such as

$$\bullet^n \boldsymbol{B}\ \forall x.((p\ x) = \bot)$$
$$\bullet^n \boldsymbol{B}\ \forall x.((q\ x) = \bot),$$

for some $n \geq 0$, in the belief theory.

Example 5.6.3. In Example 5.6.2, it was assumed that it was appropriate to maintain explicit definitions for the predicate p at the various times. If these definitions of p are large, it may be better to give the definition at each time by an incremental change to the one at the last time. This example shows how to do this and provides an interesting use of modalities applied to predicates.

Suppose that the belief theory for an agent includes the following local assumptions.

$$\boldsymbol{B}\ \forall x.((p\ x) = \ \textit{if}\ x = B\ \textit{then}\ \top\ \textit{else if}\ x = C\ \textit{then}\ \bot\ \textit{else}\ (\bullet p\ x))$$
$$\bullet\boldsymbol{B}\ \forall x.((p\ x) = \ \textit{if}\ x = A\ \textit{then}\ \top\ \textit{else if}\ x = C\ \textit{then}\ \top\ \textit{else}\ \bot).$$

Here, the definition of p at the current time (which is the same as for Example 5.6.2) is given as an incremental change to the definition of p at the last time. Note the occurrence of $\bullet p$ in the first formula in which the modality $\bullet$ is applied to the predicate p. It is also assumed that the belief theory contains the global assumption

$$\bullet\boldsymbol{B}\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}\bullet\boldsymbol{\varphi}.$$

Figure 5.32 gives the computation using $\boldsymbol{B}$ of $(p\ A)$, in which use is made of the global assumptions (5.1.2) and (5.1.4). It follows from this that $\boldsymbol{B}((p\ A) = \top)$ is a consequence of the belief theory.

If this chain of incrementally modified definitions goes back too far, it can become cumbersome. In this case, it may be better to cut the chain off at some time in the recent past. For instance, for the belief theory of this example, one can essentially recover the explicit definition of p at the current time by means of the computation of rank 1 in Figure 5.33 in which there is a substitution of the definition of p at the last time. The last step in this computation uses the following instance of the global assumption (5.1.2).

$$\bullet\varphi = \varphi,$$

$\underline{(p\ A)}$
$\mathit{if}\ \underline{A = B}\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ A = C\ \mathit{then}\ \bot\ \mathit{else}\ (\bullet p\ A)$
$\underline{\mathit{if}\ \bot\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ A = C\ \mathit{then}\ \bot\ \mathit{else}\ (\bullet p\ A)}$
$\mathit{if}\ \underline{A = C}\ \mathit{then}\ \bot\ \mathit{else}\ (\bullet p\ A)$
$\underline{\mathit{if}\ \bot\ \mathit{then}\ \bot\ \mathit{else}\ (\bullet p\ A)}$
$\underline{(\bullet p\ A)}$
$\bullet\underline{(p\ A)}$
$\underline{\bullet\ \mathit{if}\ A = A\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ A = C\ \mathit{then}\ \top\ \mathit{else}\ \bot}$
$\mathit{if}\ \underline{A = A}\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ A = C\ \mathit{then}\ \top\ \mathit{else}\ \bot$
$\underline{\mathit{if}\ \top\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ A = C\ \mathit{then}\ \top\ \mathit{else}\ \bot}$
$\top$

Figure 5.32: Computation of rank 1 using $\boldsymbol{B}$ of $(p\ A)$

$\boldsymbol{B}\ \forall x.((p\ x) = \mathit{if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ \bot\ \mathit{else}\ \underline{(\bullet p\ x)})$
$\boldsymbol{B}\ \forall x.((p\ x) = \mathit{if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ \bot\ \mathit{else}\ \bullet\underline{(p\ x)})$
$\boldsymbol{B}\ \forall x.((p\ x) = \mathit{if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ \bot$
$\quad\quad \mathit{else}\ \underline{\bullet(\mathit{if}\ x = A\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ \top\ \mathit{else}\ \bot)})$
$\boldsymbol{B}\ \forall x.((p\ x) = \mathit{if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ \bot$
$\quad\quad \mathit{else}\ \mathit{if}\ x = A\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ \top\ \mathit{else}\ \bot)$

Figure 5.33: Computation of $\boldsymbol{B}\ \forall x.((p\ x) = \mathit{if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ \bot\ \mathit{else}\ (\bullet p\ x))$

where $\varphi \triangleq \mathit{if}\ x = A\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ \top\ \mathit{else}\ \bot$ is rigid. The last formula in the computation can be simplified, if desired, to recover

$$\boldsymbol{B}\ \forall x.((p\ x) = \ \mathit{if}\ x = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = A\ \mathit{then}\ \top\ \mathit{else}\ \bot).$$

Note that the computation in Figure 5.33 does not use the leftmost selection rule. Instead the computation can be thought of as a partial evaluation of the definition of p at the current time that uses an appropriate selection of redexes and stops at an appropriate point (to produce the desired outcome from the partial evaluation).

Even though equality of the first and last formulas in Figure 5.33 has been shown to be a consequence of the (original) belief theory, a further restriction is needed to ensure their computational equivalence. This restriction is that the last formula is only ever used to compute terms of the form $(p\ t)$, where t is rigid. Figures 5.34 and 5.35 give the respective computations of $(p\ t)$ in each case. The rigidity condition is needed at several places in

the computation in Figure 5.34. Therefore, to maintain the computational equivalence of the two belief theories, it needs to be imposed in Figure 5.35. In practice, arguments to constants in belief theories are usually rigid, so this restriction is unlikely to cause any inconvenience.

$$\underline{(p\ t)}$$
$$\mathit{if}\ t = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ t = C\ \mathit{then}\ \bot\ \mathit{else}\ \underline{(\bullet p\ t)}$$
$$\mathit{if}\ t = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ t = C\ \mathit{then}\ \bot\ \mathit{else}\ \bullet\underline{(p\ t)}$$
$$\mathit{if}\ t = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ t = C\ \mathit{then}\ \bot$$
$$\qquad \mathit{else}\ \bullet\underline{(\mathit{if}\ t = A\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ t = C\ \mathit{then}\ \top\ \mathit{else}\ \bot)}$$
$$\mathit{if}\ t = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ t = C\ \mathit{then}\ \bot$$
$$\qquad \mathit{else}\ \mathit{if}\ t = A\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ t = C\ \mathit{then}\ \top\ \mathit{else}\ \bot$$

Figure 5.34: Partial computation using $\boldsymbol{B}$ of $(p\ t)$ for the original belief theory

$$\underline{(p\ t)}$$
$$\mathit{if}\ t = B\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ t = C\ \mathit{then}\ \bot$$
$$\qquad \mathit{else}\ \mathit{if}\ t = A\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ t = C\ \mathit{then}\ \top\ \mathit{else}\ \bot$$

Figure 5.35: Partial computation using $\boldsymbol{B}$ of $(p\ t)$ for new belief theory

Use of modalities applied to predicates can be avoided in Example 5.6.3 by working throughout with $\bullet(p\ x)$ instead of $(\bullet p\ x)$. However, in the next example, modal terms will become essential when the predicate p is replaced by a constant f having a signature of the form $\sigma \to \tau$ (because $(f\ x)$ will not generally be a formula).

Example 5.6.4. In Example 5.6.3, it was shown how it is possible to maintain incremental changes of definitions of a predicate at successive times. In this example, the technique is extended to constants that do not have a signature of the form $\alpha \to o$.

Suppose that the constant $f : \sigma \to \mathit{Nat}$ has definitions at the current and last times as follows.

$$\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = C\ \mathit{then}\ 3\ \mathit{else}\ (\bullet f\ x))$$
$$\bullet\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = A\ \mathit{then}\ 42\ \mathit{else}\ \mathit{if}\ x = B\ \mathit{then}\ 21\ \mathit{else}\ \mathit{if}\ x = C\ \mathit{then}\ 15\ \mathit{else}\ 0)$$

The value of f at the current time on some argument can be computed using the definition of f as in Figure 5.36. It follows from this computation that $\boldsymbol{B}((f\ A) = 42)$ is a consequence of the belief theory.

$\underline{(f\ A)}$
$\mathit{if}\ \underline{A = C}\ \mathit{then}\ 3\ \mathit{else}\ (\bullet f\ A)$
$\underline{\mathit{if}\ \bot\ \mathit{then}\ 3\ \mathit{else}\ (\bullet f\ A)}$
$\underline{(\bullet f\ A)}$
$\bullet\underline{(f\ A)}$
$\bullet\mathit{if}\ \underline{A = A}\ \mathit{then}\ 42\ \mathit{else\ if}\ A = B\ \mathit{then}\ 21\ \mathit{else\ if}\ A = C\ \mathit{then}\ 15\ \mathit{else}\ 0$
$\bullet\underline{\mathit{if}\ \top\ \mathit{then}\ 42\ \mathit{else\ if}\ A = B\ \mathit{then}\ 21\ \mathit{else\ if}\ A = C\ \mathit{then}\ 15\ \mathit{else}\ 0}$
$\bullet 42$
42

Figure 5.36: Computation using $\boldsymbol{B}$ of $(f\ A)$

The definition of f at the current time can be written equivalently as

$$\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = C\ \mathit{then}\ 3\ \mathit{else\ if}\ x = A\ \mathit{then}\ 42\ \mathit{else\ if}\ x = B\ \mathit{then}\ 21\ \mathit{else}\ 0).$$

This follows from the partial evaluation of Figure 5.37 which is a computation of rank 1 of $\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = C\ \mathit{then}\ 3\ \mathit{else}\ (\bullet f\ x))$. In a similar way to Example 5.6.3, this formula should only be used to compute terms of the form $(f\ t)$, where t is rigid.

$\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = C\ \mathit{then}\ 3\ \mathit{else}\ \underline{(\bullet f\ x)})$
$\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = C\ \mathit{then}\ 3\ \mathit{else}\ \bullet\underline{(f\ x)})$
$\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = C\ \mathit{then}\ 3$
$\quad \mathit{else}\ \bullet(\mathit{if}\ x = A\ \mathit{then}\ 42\ \mathit{else\ if}\ x = B\ \mathit{then}\ 21\ \mathit{else\ if}\ x = C\ \mathit{then}\ 15\ \mathit{else}\ 0))$
$\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = C\ \mathit{then}\ 3$
$\quad \mathit{else\ if}\ x = A\ \mathit{then}\ 42\ \mathit{else\ if}\ x = B\ \mathit{then}\ 21\ \mathit{else\ if}\ x = C\ \mathit{then}\ 15\ \mathit{else}\ 0)$

Figure 5.37: Computation of rank 1 of $\boldsymbol{B}\ \forall x.((f\ x) = \mathit{if}\ x = C\ \mathit{then}\ 3\ \mathit{else}\ (\bullet f\ x))$

Example 5.6.5. This example shows how to handle belief formulas whose body is an abstraction. Consider an agent with belief modality $\boldsymbol{B}$ which has a belief theory that includes

$$\bullet\boldsymbol{B}(f = \lambda x.\mathit{if}\ (p\ x)\ \mathit{then}\ A\ \mathit{else\ if}\ (q\ x)\ \mathit{then}\ B\ \mathit{else}\ C),$$

where A, B, and C are rigid constants, and $f : \alpha \to \beta$. Now let t be a rigid term of type α and suppose the agent wants to compute the value of $(\bullet f\ t)$. This computation may then proceed as in Figure 5.38. It follows from Figure 5.38 that $\boldsymbol{B}((\bullet f\ t) = B)$ is a consequence of the belief theory.

$(\bullet\underline{f\ t})$
$(\underline{\bullet\lambda x.\mathit{if}\ (p\ x)\ \mathit{then}\ A\ \mathit{else\ if}\ (q\ x)\ \mathit{then}\ B\ \mathit{else}\ C}\ \ t)$
$\underline{(\lambda x.\bullet\mathit{if}\ (p\ x)\ \mathit{then}\ A\ \mathit{else\ if}\ (q\ x)\ \mathit{then}\ B\ \mathit{else}\ C\ \ t)}$
$\bullet\mathit{if}\ \underline{(p\ t)}\ \mathit{then}\ A\ \mathit{else\ if}\ (q\ t)\ \mathit{then}\ B\ \mathit{else}\ C$
$\vdots$
$\bullet\underline{\mathit{if}\ \bot\ \mathit{then}\ A\ \mathit{else\ if}\ (q\ t)\ \mathit{then}\ B\ \mathit{else}\ C}$
$\bullet\mathit{if}\ \underline{(q\ t)}\ \mathit{then}\ B\ \mathit{else}\ C$
$\vdots$
$\bullet\underline{\mathit{if}\ \top\ \mathit{then}\ B\ \mathit{else}\ C}$
$\underline{\bullet B}$
B

Figure 5.38: Computation with abstractions and modalities

This computation has used the following instance of global assumption (5.1.6).

$$\bullet\lambda x.s = \lambda x.\bullet s,$$

where $s \triangleq \mathit{if}\ (p\ x)\ \mathit{then}\ A\ \mathit{else\ if}\ (q\ x)\ \mathit{then}\ B\ \mathit{else}\ C$. It has also used the global assumption

$$\bullet\boldsymbol{B}\varphi \longrightarrow \boldsymbol{B}\bullet\varphi,$$

where $\varphi \triangleq f = \lambda x.\mathit{if}\ (p\ x)\ \mathit{then}\ A\ \mathit{else\ if}\ (q\ x)\ \mathit{then}\ B\ \mathit{else}\ C$, to switch the $\bullet$ and $\boldsymbol{B}$.

Example 5.6.6. This example shows the usefulness of modal terms for predicate rewrite systems discussed in Section B.1.12.

Consider the following predicate rewrite system.

$$\mathit{top} \rightarrowtail \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \mathit{top}\ \mathit{top}))$$
$$\mathit{top} \rightarrowtail \bullet\boldsymbol{B}_j\mathit{top}$$
$$\mathit{top} \rightarrowtail \blacklozenge\boldsymbol{B}_j\mathit{top}$$
$$\mathit{top} \rightarrowtail p$$
$$\mathit{top} \rightarrowtail q$$
$$\mathit{top} \rightarrowtail r.$$

The following is a path in the predicate space defined by the rewrite system.

$$\mathit{top} \rightsquigarrow \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \mathit{top}\ \mathit{top})) \rightsquigarrow \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \bullet\boldsymbol{B}_j\mathit{top}\ \ \mathit{top}))$$
$$\rightsquigarrow \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \bullet\boldsymbol{B}_j p\ \ \mathit{top}))$$
$$\rightsquigarrow \cdots \rightsquigarrow \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \bullet\boldsymbol{B}_j p\ \blacklozenge\boldsymbol{B}_j q)).$$

Thus the predicate $\boldsymbol{B}_i(setExists_1\ (\wedge_2\ \bullet\boldsymbol{B}_j p\ \blacklozenge\boldsymbol{B}_j q))$ is in the space of predicates generated by the predicate rewrite system. If s is a set, the informal meaning of the formula

$$(\boldsymbol{B}_i(setExists_1\ (\wedge_2\ \bullet\boldsymbol{B}_j p\ \blacklozenge\boldsymbol{B}_j q))\ \ s)$$

is that agent i believes that there is an item in the set s such that at the last time step agent j believed the item satisfied p and that at some time in the past agent j believed that the item satisfied q.

Example 5.6.7. This example illustrates how modal terms might be useful for learning applications. Consider an agent i that selects an action partly in response to the perceived belief formulas of another agent j. Suppose that it matters whether the agent j believes something now or believed something at the last time or always in the past. Then a suitable predicate rewrite system for the hypothesis language associated with this learning task might be something similar to the following.

$$\begin{aligned}
&top \rightarrowtail \boldsymbol{B}_j(p\ (\wedge_2\ top\ top))\\
&top \rightarrowtail \boldsymbol{B}_j\bullet(p\ (\wedge_2\ top\ top))\\
&top \rightarrowtail \boldsymbol{B}_j\blacksquare(p\ (\wedge_2\ top\ top))\\
&top \rightarrowtail r\\
&top \rightarrowtail s\\
&top \rightarrowtail t\\
&\quad\ \vdots
\end{aligned}$$

Thus predicates such as

$$\boldsymbol{B}_j\bullet(p\ (\wedge_2\ r\ s))$$

and

$$\boldsymbol{B}_j\blacksquare(p\ (\wedge_2\ t\ r))$$

could appear in the belief theory for agent i.

Now suppose this agent computes its policy on some argument to determine which action to select. During this computation, formulas such as

$$(\boldsymbol{B}_j\bullet(p\ (\wedge_2\ r\ s))\ t),$$

for some rigid term t, will need to be evaluated. This can be done by using the global assumption (5.1.4) to turn it into

$$\boldsymbol{B}_j\bullet((p\ (\wedge_2\ r\ s))\ t).$$

Next simple cases of probabilistic reasoning are examined.

Example 5.6.8. Consider a conditional density $f : X \to \mathfrak{D}(\mathbb{B})$. Then a typical piecewise-constant definition for f might be as follows: for all $x \in X$,

$$f(x) = \begin{cases} \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.1\ \mathit{else}\ 0.9 & \text{if } p_1(x) \text{ holds} \\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.7\ \mathit{else}\ 0.3 & \text{if } p_2(x) \text{ holds but not } p_1(x) \\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.8\ \mathit{else}\ 0.2 & \text{if } p_3(x) \text{ holds but not } p_1(x) \text{ and not } p_2(x) \\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0\ \mathit{else}\ 1 & \text{otherwise.} \end{cases}$$

Note how the predicates p_1, p_2, and p_3 divide up the domain into regions on each of which the function has the same density as a value – this is typical of such functions that are acquired during deployment of an agent.

Now the logicization of f is given. Suppose the type of elements in X is σ. Recall the synonym $\mathit{Density}\ \tau \triangleq \tau \to \mathit{Real}$, where the understanding is that the meaning of a term of type $\mathit{Density}\ \tau$ is a probability density over elements of type τ rather than an arbitrary real-valued function over elements of type τ. Then the logicization of f is a constant, also denoted also by f, having signature $\sigma \to \mathit{Density\ Bool}$. The definition for f in the logic is as follows:

$$\begin{aligned} &\forall x.((f\ x) = \\ &\quad \mathit{if}\ (p_1\ x)\ \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.1\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0.9\ \mathit{else}\ 0 \\ &\quad \mathit{else\ if}\ (p_2\ x)\ \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.7\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0.3\ \mathit{else}\ 0 \\ &\quad \mathit{else\ if}\ (p_3\ x)\ \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.8\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0.2\ \mathit{else}\ 0 \\ &\quad \mathit{else}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 1\ \mathit{else}\ 0). \end{aligned}$$

It will be convenient because of the operations that will be applied to *if_then_else* expressions to be a little pedantic about how they are written by explicitly having an '*else* 0' case at the end. Note carefully the use that has been made of the higher-orderness of the logic; this approach is not available with first-order logic.

In preparation for Example 5.6.11 below, the definitions of the functions ♮, ♭, §, and \$ that are special cases of the fusion $\odot$ of conditional densities given in Definition A.3.7 are now presented. For this, attention is confined in the following to the discrete case for which densities have finite support (that is, take a non-zero value on at most finitely many elements of their domain). For this case, several summation functions whose arguments are functions with finite support are needed.

First comes the constant *sum1* that computes $\sum_x f(x)$, where f is a real-valued function having finite support (that is, takes a non-zero value on at most finitely many elements of its domain) and x ranges over the domain of f.

$$\begin{aligned} &\mathit{sum1} : (a \to \mathit{Real}) \to \mathit{Real} \\ &(\mathit{sum1}\ \lambda x.\mathit{if}\ x = u\ \mathit{then}\ v\ \mathit{else}\ \boldsymbol{w}) = v + (\mathit{sum1}\ \lambda x.\boldsymbol{w}) \\ &(\mathit{sum1}\ \lambda x.0) = 0 \end{aligned}$$

Here, it is assumed that the argument to *sum1* has the syntactic form of an abstraction over a nested *if_then_else* for which there are no repeated occurrences of the tests $x = u$. If there are such repeats, it is easy enough to remove them.

Example 5.6.9. Figure 5.39 gives the computation of

$$(sum1\ \lambda y.if\ y = \top\ then\ 0.1\ else\ if\ y = \bot\ then\ 0.5\ else\ 0),$$

which has the answer 0.6.

$\underline{(sum1\ \lambda y.if\ y = \top\ then\ 0.1\ else\ if\ y = \bot\ then\ 0.5\ else\ 0)}$
$0.1 + \underline{(sum1\ \lambda y.if\ y = \bot\ then\ 0.5\ else\ 0)}$
$0.1 + (0.5 + \underline{(sum1\ \lambda y.0)})$
$0.1 + \underline{(0.5 + 0)}$
$\underline{0.1 + 0.5}$
0.6

Figure 5.39: Computation of $(sum1\ \lambda y.if\ y = \top\ then\ 0.1\ else\ if\ y = \bot\ then\ 0.5\ else\ 0)$

Also needed will be the constant *sum2* that computes $\sum_{x \in S} f(x)$, where f is a real-valued function having finite support and S is a subset of the domain of f.

$$sum2 : \{a\} \to (a \to Real) \to Real$$
$$(sum2\ s\ \lambda x.if\ x = u\ then\ v\ else\ \boldsymbol{w}) = (if\ u \in s\ then\ v\ else\ 0) + (sum2\ s\ \lambda x.\boldsymbol{w})$$
$$(sum2\ s\ \lambda x.0) = 0$$

Of course, '$u \in s$' is just '$(s\ u)$'. Note that *sum1* and *sum2* are closely related in that

$$(sum2\ \lambda y.\top\ f) = (sum1\ f).$$

Example 5.6.10. Figure 5.40 gives the computation of

$$(sum2\ \lambda x.(x = \bot)\ \lambda y.if\ y = \top\ then\ 0.1\ else\ if\ y = \bot\ then\ 0.5\ else\ 0),$$

which has the answer 0.5.

Consider now the following four functions.

$$\natural : (X \to \mathcal{D}(Y)) \to (Y \to \mathcal{D}(Z)) \to (X \to \mathcal{D}(Z))$$
$$(f \natural g)(x)(z) = \sum_{y \in Y} f(x)(y) \times g(y)(z), \text{ for all } x \in X \text{ and } z \in Z.$$

$$\flat : (X \to \mathcal{D}(Y)) \to (Y \to Z) \to (X \to \mathcal{D}(Z))$$
$$(f \flat g)(x)(z) = \sum_{y \in g^{-1}(z)} f(x)(y), \text{ for all } x \in X \text{ and } z \in Z.$$

$$\S : \mathcal{D}(Y) \to (Y \to \mathcal{D}(Z)) \to \mathcal{D}(Z)$$

$\underline{(sum2\ \lambda x.(x = \bot)\ \lambda y.if\ y = \top\ then\ 0.1\ else\ if\ y = \bot\ then\ 0.5\ else\ 0)}$
$if\ \underline{\top \in \lambda x.(x = \bot)}\ then\ 0.1\ else\ 0 + (sum2\ \lambda x.(x = \bot)\ \lambda y.if\ y = \bot\ then\ 0.5\ else\ 0)$
$if\ \underline{\top = \bot}\ then\ 0.1\ else\ 0 + (sum2\ \lambda x.(x = \bot)\ \lambda y.if\ y = \bot\ then\ 0.5\ else\ 0)$
$\underline{if\ \bot\ then\ 0.1\ else\ 0} + (sum2\ \lambda x.(x = \bot)\ \lambda y.if\ y = \bot\ then\ 0.5\ else\ 0)$
$\underline{0 + (sum2\ \lambda x.(x = \bot)\ \lambda y.if\ y = \bot\ then\ 0.5\ else\ 0)}$
$\underline{(sum2\ \lambda x.(x = \bot)\ \lambda y.if\ y = \bot\ then\ 0.5\ else\ 0)}$
$\vdots$
$0.5 + \underline{(sum2\ \lambda y.0)}$
$\underline{0.5 + 0}$
0.5

Figure 5.40: Computation of $(sum2\ \lambda x.(x = \bot)\ \lambda y.if\ y = \top\ then\ 0.1\ else\ if\ y = \bot\ then\ 0.5\ else\ 0)$

$$(f\ \S\ g)(z) = \sum_{y \in Y} f(y) \times g(y)(z),\ \text{for all } z \in Z.$$

$$\$: \mathcal{D}(Y) \to (Y \to Z) \to \mathcal{D}(Z)$$
$$(f\ \$\ g)(z) = \sum_{y \in g^{-1}(z)} f(y),\ \text{for all } z \in Z.$$

The intuitive meaning of the function $\natural$ is that it is the fusion of two functions each of which has uncertainty in its values giving a fused function that has the accumulated uncertainty of these two functions in its values; the other functions have similar meanings that are essentially special cases of this. Note that each of these functions is a variation of the standard fusion function $\odot$, where $h_1 : X_0 \to \mathcal{D}(X_1)$ and $h_2 : X_0 \times X_1 \to \mathcal{D}(X_2)$ are conditional densities and $h_1 \odot h_2 : X_0 \to \mathcal{D}(X_2)$ is defined by

$$(h_1 \odot h_2)(x_0) = \lambda x_2. \int_{X_1} \lambda x_1.h_2(x_0, x_1)(x_2)\ h_1(x_0)\ d\nu_1,$$

for all $x_0 \in X_0$.

Here are the corresponding definitions for the logicization of these functions.

$$\natural : (a \to Density\ b) \to (b \to Density\ c) \to (a \to Density\ c)$$
$$f \natural g = \lambda x.\lambda z.(sum1\ \lambda y.(((f\ x)\ y) \times ((g\ y)\ z)))$$

$$\flat : (a \to Density\ b) \to (b \to c) \to (a \to Density\ c)$$
$$f \flat g = \lambda x.\lambda z.(sum2\ \lambda y.((g\ y) = z)\ \lambda y.((f\ x)\ y))$$

$$\S : \mathit{Density}\ b \to (b \to \mathit{Density}\ c) \to \mathit{Density}\ c$$
$$f\ \S\ g = \lambda z.(\mathit{sum1}\ \lambda y.((f\ y) \times ((g\ y)\ z)))$$

$$\$: \mathit{Density}\ b \to (b \to c) \to \mathit{Density}\ c$$
$$f\ \$\ g = \lambda z.(\mathit{sum2}\ \lambda y.((g\ y) = z)\ f).$$

Next is an example using the functions § and $ that shows how an agent can logically reason with a mixture of certain and uncertain assumptions expressed in theories, and how conclusions obtained by reasoning with such theories can have a level of certainty directly associated with them.

Example 5.6.11. Consider an agent that makes recommendations of TV programs to a user. The agent has access to the TV guide through the definition of the function *tv_guide*. It also knows about the user's preferences for TV programs through the definition of the function *likes*, the uncertainty in which is modelled by densities in its codomain. Suppose now that the agent is asked to make a recommendation about a program at a particular occurrence (that is, date, time, and channel), except that there is some uncertainty in the actual occurrence intended by the user; this uncertainty, which is modelled in the definition of the density *choice*, could come about, for example, if the user asked the somewhat ambiguous question "Are there any good programs on *ABC* during dinner time?". The question the agent needs to answer is the following: given the uncertainty in *choice* and *likes*, what is the probability that the user likes the program that is on at the occurrence intended by the user.

This situation is modelled as follows. First, here are some type synonyms.

$$\mathit{Occurrence} = \mathit{Date} \times \mathit{Time} \times \mathit{Channel}$$
$$\mathit{Date} = \mathit{Day} \times \mathit{Month} \times \mathit{Year}$$
$$\mathit{Time} = \mathit{Hour} \times \mathit{Minute}$$
$$\mathit{Program} = \mathit{Title} \times \mathit{Duration}\ \times \mathit{Genre} \times \mathit{Classification} \times \mathit{Synopsis}.$$

Here is the definition of the constant *choice* that models the uncertainty in the intended occurrence, where $\boldsymbol{B}$ is the belief modality of the agent.

$$\mathit{choice} : \mathit{Density}\ \mathit{Occurrence}$$
$$\boldsymbol{B}\ \forall x.((\mathit{choice}\ x) =$$
$$\quad \mathit{if}\ x = ((21,7,2007),(19,30),\mathit{ABC})\ \mathit{then}\ 0.8$$
$$\quad \mathit{else\ if}\ x = ((21,7,2007),(20,30),\mathit{ABC})\ \mathit{then}\ 0.2$$
$$\quad \mathit{else}\ 0).$$

So the uncertainty is in the time of the program; it is probably 7.30pm, but it could be 8.30pm. Note that an equivalent form of the definition for *choice* is

$$\boldsymbol{B}(\mathit{choice} = \lambda x.\mathit{if}\ x = ((21,7,2007),(19,30),\mathit{ABC})\ \mathit{then}\ 0.8\ \mathit{else}\ \ldots),$$

which is the form of the definition that is used below.

Next there is the TV guide that maps occurrences to programs.

$$
\begin{array}{l}
\mathit{tv\_guide} : \mathit{Occurrence} \to \mathit{Program} \\
\boldsymbol{B}\ \forall x.((\mathit{tv\_guide}\ x) = \\
\quad \mathit{if}\ ((= ((20,7,2007),(11,30),\mathit{Win}))\ x) \\
\qquad\qquad \mathit{then}\ (\text{“}\mathit{NFL\ Football}\text{”}, 60, \mathit{Sport}, G, \text{“}\mathit{The\ Buffalo}\ldots\text{”}) \\
\quad \mathit{else\ if}\ ((= ((21,7,2007),(19,30),\mathit{ABC}))\ x) \\
\qquad\qquad \mathit{then}\ (\text{“}\mathit{Seinfeld}\text{”}, 30, \mathit{Sitcom}, \mathit{PG}, \text{“}\mathit{Kramer}\ldots\text{”}) \\
\quad \mathit{else\ if}\ ((= ((21,7,2007),(20,30),\mathit{ABC}))\ x) \\
\qquad\qquad \mathit{then}\ (\text{“}\mathit{The\ Bill}\text{”}, 50, \mathit{Drama}, M, \text{“}\mathit{Sun\ Hill}\ldots\text{”}) \\
\quad \mathit{else\ if}\ ((= ((21,7,2007),(21,30),\mathit{ABC}))\ x) \\
\qquad\qquad \mathit{then}\ (\text{“}\mathit{Numb3rs}\text{”}, 60, \mathit{Crime}, M, \text{“}\mathit{When}\ldots\text{”}) \\
\qquad \vdots \\
\quad \mathit{else}\ (\text{“ ”}, 0, \mathit{NULL}, \mathit{NA}, \text{“ ”})).
\end{array}
$$

Finally, here is the definition of the constant *likes*. (A more realistic modelling of the situation would involve also using the belief modality of the user in the definition of *likes*, but this is left out for simplicity.)

$$
\begin{array}{l}
\mathit{likes} : \mathit{Program} \to \mathit{Density\ Bool} \\
\boldsymbol{B}\ \forall x.((\mathit{likes}\ x) = \\
\quad \mathit{if}\ (\mathit{projTitle} \diamond (=\ \text{“}\mathit{NFL\ Football}\text{”})\ x) \\
\qquad\qquad \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 1\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0\ \mathit{else}\ 0 \\
\quad \mathit{else\ if}\ (\mathit{projGenre} \diamond (=\ \mathit{Movie})\ x) \\
\qquad\qquad \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.75\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0.25\ \mathit{else}\ 0 \\
\quad \mathit{else\ if}\ (\mathit{projGenre} \diamond (=\ \mathit{Documentary})\ x) \\
\qquad\qquad \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 1\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0\ \mathit{else}\ 0 \\
\quad \mathit{else}\ (\bullet \mathit{likes}\ x)) \\
\bullet\boldsymbol{B}\ \forall x.((\mathit{likes}\ x) = \\
\quad \mathit{if}\ (\mathit{projTitle} \diamond (=\ \text{“}\mathit{NFL\ Football}\text{”})\ x) \\
\qquad\qquad \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.9\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0.1\ \mathit{else}\ 0 \\
\quad \mathit{else\ if}\ (\mathit{projGenre} \diamond (=\ \mathit{Drama})\ x) \\
\qquad\qquad \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.2\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0.8\ \mathit{else}\ 0 \\
\quad \mathit{else}\ (\bullet \mathit{likes}\ x)) \\
\bullet^2\boldsymbol{B}\ \forall x.((\mathit{likes}\ x) = \\
\quad \mathit{if}\ (\mathit{projGenre} \diamond (=\ \mathit{Sitcom})\ x) \\
\qquad\qquad \mathit{then}\ \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.9\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0.1\ \mathit{else}\ 0 \\
\quad \mathit{else}\ (\bullet \mathit{likes}\ x)) \\
\bullet^3\boldsymbol{B}\ \forall x.((\mathit{likes}\ x) = \\
\quad \lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 1\ \mathit{else}\ 0).
\end{array}
$$

Recall that the agent needs to compute the probability that the user likes the program which is on at the occurrence intended by the user. This amounts to computing (using $\boldsymbol{B}$) the value of the term

$$((\mathit{choice}\ \$\ \mathit{tv\_guide})\ \S\ \mathit{likes}),$$

which is a term of type *Density Bool*. The answer for this computation is

$$\lambda z.\mathit{if}\ z = \top\ \mathit{then}\ 0.76\ \mathit{else}\ \mathit{if}\ z = \bot\ \mathit{then}\ 0.24\ \mathit{else}\ \mathit{if}\ z = \top\ \mathit{then}\ 0.72\ \mathit{else}\ \mathit{if}\ z = \bot\ \mathit{then}\ 0.08\ \mathit{else}\ 0,$$

which can be simplified to

$$\lambda z.\mathit{if}\ z = \top\ \mathit{then}\ 0.76\ \mathit{else}\ \mathit{if}\ z = \bot\ \mathit{then}\ 0.24\ \mathit{else}\ 0.$$

Thus

$$\boldsymbol{B}(((\mathit{choice}\ \$\ \mathit{tv\_guide})\ \S\ \mathit{likes}) = \lambda z.\mathit{if}\ z = \top\ \mathit{then}\ 0.76\ \mathit{else}\ \mathit{if}\ z = \bot\ \mathit{then}\ 0.24\ \mathit{else}\ 0)$$

is a consequence of the belief theory of the agent. On this basis, at least in part, the agent can now decide whether to recommend the program or not.

The computation is somewhat complicated. To make it easier to understand, the computation is broken up into two steps.

The computation of $(\mathit{choice}\ \$\ \mathit{tv\_guide})$, shown in Figure 5.41, gives the answer

$$\lambda y.\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, 30, \mathit{Sitcom}, \mathit{PG}, \text{“}\mathit{Kramer}\ldots\text{”})\ \mathit{then}\ 0.8\ \mathit{else}\ \mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, 50, \mathit{Drama}, M, \text{“}\mathit{Sun\ Hill}\ldots\text{”})\ \mathit{then}\ 0.2\ \mathit{else}\ 0,$$

which is a term of type *Density Program*. Then the second part of the computation in Figure 5.42 is that of

$$((\lambda y.\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8\ \ \mathit{else}\ \mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0)\ \S\ \mathit{likes})$$

that has the answer

$$\lambda z.\mathit{if}\ z = \top\ \mathit{then}\ 0.76\ \mathit{else}\ \mathit{if}\ z = \bot\ \mathit{then}\ 0.24\ \mathit{else}\ \mathit{if}\ z = \top\ \mathit{then}\ 0.72\ \mathit{else}\ \mathit{if}\ z = \bot\ \mathit{then}\ 0.08\ \mathit{else}\ 0.$$

The preceding example illustrates that higher-order logic is an excellent setting for integrating probability and logic. Indeed, in this setting, no new conceptual machinery is needed at all to achieve the integration. Probabilistic constructs, such as densities, can be naturally included in higher-order theories. Furthermore, the connectives, quantifiers, modalities, and probabilistic constructs all fit harmoniously together.

5.7 Reasoning about Empirical Beliefs

To do: Up to now, the reasoning system has been applied to mostly generic reasoning tasks. Now it is time to reason about the kinds of empirical beliefs that can be acquired by the methods of Chapter 4. This means that special attention must be paid to the codomains of such beliefs which can contain arbitrary distributions, including all the standard distributions that are widely used in applications, such as Gaussian, Dirichlet, categorical, multinomial, and so on. The implementation of the reasoning system will need specialized support for this.

$$
\begin{aligned}
&\underline{(\mathit{choice}\ \$\ \mathit{tv\_guide})}\\
&\lambda z.(\mathit{sum2}\ \lambda y.(\underline{(\mathit{tv\_guide}\ y)} = z)\ \mathit{choice})\\
&\lambda z.(\mathit{sum2}\ \lambda y.((\mathit{if}\ y = ((20,7,2007),(11,30),\mathit{Win})\ \mathit{then}\ \ldots) = z)\ \underline{\mathit{choice}})\\
&\lambda z.(\underline{\mathit{sum2}\ \lambda y.((\mathit{if}\ y = ((20,7,2007),(11,30),\mathit{Win})\ \mathit{then}\ \ldots) = z)}\\
&\qquad\qquad\qquad\qquad\underline{\lambda x.\mathit{if}\ x = ((21,7,2007),(19,30),\mathit{ABC})\ \mathit{then}\ 0.8\ \ldots})\\
&\quad\vdots\\
&\lambda z.(\mathit{if}\ z = (\text{“}\mathit{Seinfeld}\text{”},\ldots)\ \mathit{then}\ 0.8\ \mathit{else}\ 0\\
&\quad + (\underline{\mathit{sum2}\ \lambda y.((\mathit{if}\ y = ((20,7,2007),(11,30),\mathit{Win})\ \mathit{then}\ \ldots) = z)}\\
&\qquad\qquad\qquad\qquad\underline{\lambda x.\mathit{if}\ x = ((21,7,2007),(20,30),\mathit{ABC})\ \mathit{then}\ 0.2\ \ldots})\\
&\quad\vdots\\
&\lambda z.(\mathit{if}\ z = (\text{“}\mathit{Seinfeld}\text{”},\ldots)\ \mathit{then}\ 0.8\ \mathit{else}\ 0\\
&\quad + \mathit{if}\ z = (\text{“}\mathit{The\ Bill}\text{”},\ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0\\
&\quad + \underline{(\mathit{sum2}\ \lambda y.((\mathit{if}\ y = ((20,7,2007),(11,30),\mathit{Win})\ \mathit{then}\ \ldots) = z)\ \lambda x.0)}\\
&\lambda z.(\mathit{if}\ z = (\text{“}\mathit{Seinfeld}\text{”},\ldots)\ \mathit{then}\ 0.8\ \mathit{else}\ 0\\
&\quad + \underline{\mathit{if}\ z = (\text{“}\mathit{The\ Bill}\text{”},\ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0\ +\ 0)}\\
&\lambda z.(\underline{\mathit{if}\ z = (\text{“}\mathit{Seinfeld}\text{”},\ldots)\ \mathit{then}\ 0.8\ \mathit{else}\ 0\ +\ \mathit{if}\ z = (\text{“}\mathit{The\ Bill}\text{”},\ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0})\\
&\quad\vdots\\
&\lambda z.\mathit{if}\ z = (\text{“}\mathit{Seinfeld}\text{”},\ldots)\ \mathit{then}\ 0.8\ \mathit{else}\ \mathit{if}\ z = (\text{“}\mathit{The\ Bill}\text{”},\ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0
\end{aligned}
$$

Figure 5.41: Computation of $(\mathit{choice}\ \$\ \mathit{tv\_guide})$

5.8 Reasoning for Choosing Actions

Now that the details of reasoning with beliefs have been presented, the discussion turns to the more general topic of the use of (empirical) beliefs for acting and communicating.

First, it is necessary to consider what an empirical belief base for an agent in a typical application might look like. Typically, applications have an environment that consists of a number of 'actors', where an 'actor' may be a person, robot, autonomous vehicle, or similar, that interact on some 'stage' that may be a room, a building, a road system, or similar. Generally, it makes sense to model the 'stage' separately from the 'actors'. Suppose the 'stage' can be modelled as a state S, in the sense of Section 2. The corresponding empirical belief of the agent about the state at some time step thus has a signature $\mathcal{P}(S)$. The agent will also have some empirical beliefs about each of the 'actors' in the environment. These typically have signatures of the form $\mathcal{P}(W^Z)$, for some W and Z. Each of these empirical beliefs may require some deconstruction. Then the beliefs, or those obtained from any deconstruction, are logicized as belief formulas and added to the belief theory. In this form, they are available for a number of uses.

$$\underline{((\lambda y.\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8\ \ \mathit{else\ if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0) \S \mathit{likes})}$$

$$\lambda z.(\mathit{sum1}\ \lambda y.(\underline{((\lambda y.\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8\ \mathit{else}}$$
$$\underline{\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0)\ y)} \times ((\mathit{likes}\ y)\ z)))$$

$$\lambda z.(\mathit{sum1}\ \lambda y.(\underline{(\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8\ \mathit{else}}$$
$$\underline{\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0)} \times ((\mathit{likes}\ y)\ z)))$$

$$\vdots$$

$$\lambda z.(\mathit{sum1}\ \lambda y.(\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8 \times (\underline{(\mathit{likes}\ (\text{“}\mathit{Seinfeld}\text{”}, \ldots))}\ z)$$
$$\mathit{else}\ ((\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0) \times ((\mathit{likes}\ y)\ z))))$$

$$\vdots$$

$$\lambda z.(\mathit{sum1}\ \lambda y.(\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8 \times (\underline{(\bullet \mathit{likes}\ (\text{“}\mathit{Seinfeld}\text{”}, \ldots))}\ z)$$
$$\mathit{else}\ ((\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0) \times ((\mathit{likes}\ y)\ z))))$$

$$\vdots$$

$$\lambda z.(\mathit{sum1}\ \lambda y.(\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8 \times (\underline{(\bullet^2 \mathit{likes}\ (\text{“}\mathit{Seinfeld}\text{”}, \ldots))}\ z)$$
$$\mathit{else}\ ((\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0) \times ((\mathit{likes}\ y)\ z))))$$

$$\vdots$$

$$\lambda z.(\mathit{sum1}\ \lambda y.(\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8 \times \underline{(\lambda y.\mathit{if}\ y = \top\ \mathit{then}\ 0.9\ \mathit{else\ if}\ y = \bot\ \mathit{then}\ 0.1\ \mathit{else}\ 0\ \ z)}$$
$$\mathit{else}\ ((\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0) \times ((\mathit{likes}\ y)\ z))))$$

$$\lambda z.(\mathit{sum1}\ \lambda y.(\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ \underline{0.8 \times (\mathit{if}\ z = \top\ \mathit{then}\ 0.9\ \mathit{else\ if}\ z = \bot\ \mathit{then}\ 0.1\ \mathit{else}\ 0)}$$
$$\mathit{else}\ ((\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0) \times ((\mathit{likes}\ y)\ z))))$$

$$\vdots$$

$$\lambda z.(\mathit{sum1}\ \lambda y.(\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ (\mathit{if}\ z = \top\ \mathit{then}\ 0.72\ \mathit{else\ if}\ z = \bot\ \mathit{then}\ 0.08\ \mathit{else}\ 0)\ \mathit{else}$$
$$\underline{((\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ 0.2\ \mathit{else}\ 0) \times ((\mathit{likes}\ y)\ z))}))$$

$$\vdots$$

$$\lambda z.\underline{(\mathit{sum1}\ \lambda y.(\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ (\mathit{if}\ z = \top\ \mathit{then}\ 0.72\ \mathit{else\ if}\ z = \bot\ \mathit{then}\ 0.08\ \mathit{else}\ 0)\ \mathit{else}}$$
$$\underline{\mathit{if}\ y = (\text{“}\mathit{The\ Bill}\text{”}, \ldots)\ \mathit{then}\ (\mathit{if}\ z = \top\ \mathit{then}\ 0.04\ \mathit{else\ if}\ z = \bot\ \mathit{then}\ 0.16\ \mathit{else}\ 0)\ \mathit{else}\ 0))}$$

$$\vdots$$

$$\lambda z.(\underline{(\mathit{if}\ z = \top\ \mathit{then}\ 0.72\ \mathit{else\ if}\ z = \bot\ \mathit{then}\ 0.08\ \mathit{else}\ 0) +}$$
$$\underline{(\mathit{if}\ z = \top\ \mathit{then}\ 0.04\ \mathit{else\ if}\ z = \bot\ \mathit{then}\ 0.16\ \mathit{else}\ 0))}$$

$$\vdots$$

$$\lambda z.\mathit{if}\ z = \top\ \mathit{then}\ 0.76\ \mathit{else\ if}\ z = \bot\ \mathit{then}\ 0.24\ \mathit{else\ if}\ z = \top\ \mathit{then}\ 0.72\ \mathit{else\ if}\ z = \bot\ \mathit{then}\ 0.08\ \mathit{else}\ 0$$

Figure 5.42: Computation of $((\lambda y.\mathit{if}\ y = (\text{“}\mathit{Seinfeld}\text{”}, \ldots)\ \mathit{then}\ 0.8\ \mathit{else}\ \ldots) \S \mathit{likes})$

To do: This is where everything has to be brought together! Chapter 4 shows how stochastic filtering, and the special case of Bayesian inference, can be employed to acquire empirical beliefs. This means that, in particular, all the methods of Bayesian machine learning, including Bayesian deep learning, can be brought to bear to acquire empirical beliefs. These beliefs have the form of conditional densities or probability kernels; thus their codomains are spaces of distributions. The empirical beliefs are then logicized so that they can be reasoned about. The logic of the reasoning system, modal higher-order logic, is a highly expresssive logic that can model in great detail the probabilistic and modal properties of empirical beliefs. Overall, this is a straightforward and effective way to combine the machine learning and logical reasoning branches of the field of artificial intelligence. Now the primary purpose of an agent is to act in such a way as to achieve its goals; in this context, acting includes communicating with humans and other agents. As a key part of deciding how to act, which may be via, say, reinforcement learning, the agent will need to reason about its beliefs. This section concentrates on such reasoning tasks. Note that reasoning includes computation such as may be carried out by a functional programming language. The aim will be to give applications that show convincingly that the agent architecture proposed here supports well these reasoning tasks.

Bibliographical Notes

Useful references on higher-order logic include [6, 26, 93, 96, 142, 158]. For a highly readable account of the advantages of working in higher-order logic rather than first-order, [48] is recommended. For modal higher-order logic, see, for example, [11, 50, 97, 116].

The standard way of modelling mentalistic concepts such as knowledge, beliefs, intentions, and so on, is with modal logic and since there are a number of such concepts and generally a number of agents in any application, one is led to the need for a multi-modal logic. While modal *propositional* logics are commonly used to analyse agents (see, for example, [47, 53, 105, 172]), to adequately model agent beliefs, the logic must be much more expressive than propositional logic. The need for modal *higher-order* logic is argued in this book. Earlier work on the approach to the representation and acquisition of beliefs that this book is based upon can be found in [98, 99, 100, 120].

The reasoning system employed in this book is based on the programing language Bach [101]. The design of Bach continues one thread in the development of declarative programming languages that goes back over 20 years. The starting point was the recognition that Prolog has various flaws that reduce its credibility as a declarative programming language; these include non-declarative meta-programming facilities and the lack of a type system [94]. This motivated the Gödel programming language [75] that was closely based on Prolog but had a polymorphic type system and declarative meta-programming facilities. The next step was Escher [95] that differed markedly from Gödel in that it was a higher-order language and was based on equational theories rather than clausal theories. In its final form, Escher was presented as an extension to Haskell [130], thus taking advantage of the many good design decisions of that language, by adding the idea of programming with abstractions [96] that provides the logic programming idioms. Escher also avoided the highly problematical negation as failure rule by treating negation as just another function. Bach builds on Escher mainly by providing modal and probabilistic computation

that is especially useful for agent applications. The extension to probabilistic theories requires no extension of the logic since higher-order functions are sufficient to represent and reason about probability densities, although *efficient* probabilistic reasoning does require additional support at the programming language level.

The material on probabilistic reasoning in this book is set in the context of the general problem of integrating logic and probability, a long-standing problem that is currently attracting substantial interest. There are three main threads to the study of the integration of logic and probability. The oldest by far is the philosophical thread that can be traced via Boole [18, 19] back to Jacob Bernoulli in 1713. An extensive historical account of this thread can be found in [64] and overviews of more recent work in [37, 65, 171]. The second thread is that of the knowledge representation and reasoning community in artificial intelligence, of which [46, 57, 66, 67, 121, 122, 124, 132, 137, 143] are typical works. The third thread is that of the machine learning community in artificial intelligence, of which [36, 79, 86, 106, 107, 112, 135] are typical works.

Intuitively, the problem of integrating logic and probability is to find some way of effectively doing probabilistic reasoning in a logical formalism that may involve the invention of 'probabilistic logics'. So, for example, it is desirable that there be some way of representing the uncertainty of assumptions in a theory and also methods of computing the uncertainty of theorems that are proved from such assumptions. Here is a brief discussion of some approaches to integrating logic and probability that have come recently from the machine learning community.

Using a density to model the uncertain value of a function on some argument is better than using a single value (such as the mean of the density). For example, if the density is a normal distribution, then it may be important that the variance is large or small: if it is large, intuitively, there is less confidence about its actual value; if it is small, then there could be confidence that the actual value is the mean. Such subtleties can assist in the selection of one action over another. Similarly, learning tasks can exploit the existence of the density by including features based on the mean, variance, higher moments, or other parameters of the density in hypothesis languages.

The standard logical setting for these approaches is first-order logic. Imagine that an agent is operating in some environment for which there is some uncertainty (for example, the environment might be partially observable). The environment is modelled as a probability distribution over the collection of first-order interpretations (over some suitable alphabet for the application at hand). The intuition is that any of these interpretations could be the actual environment but that some interpretations are more likely than others to correctly model the actual world and this information is given by the distribution on the interpretations. If the agent actually knew this distribution, then it could answer probabilistic questions of the form: if (closed) formula ψ holds, what is the probability that the (closed) formula φ holds? In symbols, the question is: what is $Pr(\varphi \,|\, \psi)$?

This situation is formalized as follows. Let $\mathcal{I}$ be the set of interpretations and p a probability measure on the σ-algebra of all subsets of this set. Define the random variable $X_\varphi : \mathcal{I} \to \mathbb{R}$ by

$$X_\varphi(I) = \begin{cases} 1 & \text{if } \varphi \text{ is true in } I \\ 0 & \text{otherwise,} \end{cases}$$

with a similar definition for X_ψ. Then $Pr(\varphi \,|\, \psi)$ can be written in the form

$$p(X_\varphi = 1 \,|\, X_\psi = 1)$$

which is equal to

$$\frac{p(X_\varphi = 1 \wedge X_\psi = 1)}{p(X_\psi = 1)}$$

and, knowing p, can be evaluated.

Of course, the real problem is to know the distribution on the interpretations. To make some progress on this, most systems intending to integrate logical and probabilistic reasoning in artificial intelligence make simplifying assumptions. For a start, most are based on Prolog. Thus theories are first-order Horn clause theories, maybe with negation as failure. Interpretations are limited to Herbrand interpretations and often function symbols are excluded so the Herbrand base (and therefore the number of Herbrand interpretations) is finite. Let $\mathcal{I}$ denote the (finite) set of Herbrand interpretations and $\mathcal{B}$ the Herbrand base. One can identify $\mathcal{I}$ with the product space $\{0,1\}^{\mathcal{B}}$ in the natural way. Thus the problem amounts to knowing the distribution on this product space. At this point, there is a wide divergence in the approaches. For example, either Bayesian networks or Markov random fields can be used to represent the product distribution. The occurrences of atoms in the same clause can be used to give the edges and the weights attached to clauses are used to give the potential functions in a Markov random field. Conditional probability distributions can be attached to clauses to give a Bayesian network. Alternatively, a program is written that specifies a generative distribution for a Bayesian network. In all cases, the logic is exploited to give some kind of compact representation of what is usually a very large graphical model. Generally, the theory is only used to construct the graphical model and reasoning proceeds probabilistically, as described above.

In this book, a different approach has been followed. To begin with, a much more expressive logic – modal higher-order logic – is used. The higher-orderness is essential to achieve the desired integration of logic and probability. Also, the modalities are important for agent applications. Furthermore, in the approach here, the theory plays a central role and all probabilistic reasoning takes place in the context of the theory.

The literature on integrating logic and probability contains a variety of probabilistic logics. The syntactic features of two of these logics are now examined and it is shown how they can be easily expressed using higher-order logic. In the logic of [46], it is possible to write expressions such as $w(\varphi) < 1/3$ and $w(\varphi) \geq 2w(\psi)$, where the first means that "φ has probability less than 1/3" and the second means "φ is at least twice as probable as ψ". To write such expressions in higher-order logic, one replaces the formula φ in the logic of [46] by a density on the booleans denoted by φ'. (Some of) the function symbols in φ would have definitions involving densities. Then $w(\varphi) < 1/3$ is written in the logic of this book as

$$(\varphi'\ \top) < 1/3$$

and $w(\varphi) \geq 2w(\psi)$ is written as

$$(\varphi'\ \top) \geq 2(\psi'\ \top).$$

The notation in [46] is extended so that one can write $w_i(\varphi) < 1/3$ with the meaning "according to agent i, φ holds with probability less than 1/3". In the logic of this book, this can be expressed by

$$(\boldsymbol{B}_i\varphi'\ \top) < 1/3$$

or, equivalently by Equation 5.1.4 since $\top$ is rigid,

$$\boldsymbol{B}_i(\varphi'\ \top) < 1/3.$$

(An alternative to using $\boldsymbol{B}_i$ in the preceding two formulas is to use $\boldsymbol{K}_i$.) By the way, note carefully the difference in meaning between $\boldsymbol{B}_i(\varphi'\ \top) < 1/3$ and

$$\boldsymbol{B}_i((\varphi'\ \top) < 1/3).$$

Easy examples show that there are interpretations for which the first formula is true whereas the second is false. Even formulas such as $a_1 w_i(\varphi_1) + \cdots + a_k w_i(\varphi_k) \geq b$, where $a_1, \ldots, a_k, b$ are numbers, are admitted in [46]. This can be written as

$$a_1\boldsymbol{B}_i(\varphi_1'\ \top) + \cdots + a_k\boldsymbol{B}_i(\varphi_k'\ \top) \geq b$$

in the logic of this book.

In the logic of [66], expressions such as $w_x(\varphi) \geq 1/2$ are admitted, where this expression has the meaning that "the probability that a randomly chosen x in the domain satisfies φ is greater than or equal to 1/2". Let h be the relevant density (with finite support) and S the set $\{x \,|\, \varphi\}$, Then, mathematically, the required expression is $\sum_{x\in S} h(x) \geq 1/2$, where $\sum_{x\in S} h(x)$ is the measure of the set of x that satisfy φ. In the logic of this book, this is expressed as

$$(\mathit{sum2}\ \{x \,|\, \varphi\}\ h) \geq 1/2.$$

This chapter has illustrated the Bach programming language that follows a long tradition of declarative programming languages. The central claim is that reasoning, as described here, covers a significant part of the computation that is needed for artificial intelligence. Depending on the application, one could expect a significant proportion of the computation of an agent to be reasoning about its beliefs to decide what action to perform. If it is indeed true that reasoning is a core part of AI computation, it makes sense to ask if computers, designed specifically to do reasoning as efficiently as possible, should be used for AI applications. Indeed, this idea is an old one, going back to the Lisp machines and Prolog machines of the 1980s. For example, the Japanese Fifth Generation Project [169], which ran from 1982 to 1992, had a goal of building parallel Prolog machines. But Lisp machines have had very little impact on the computer industry and Prolog machines have had no impact at all. General-purpose computers improved so rapidy that emulating a Prolog machine on a general-purpose computer was consistently faster than running the Prolog machine itself. What triumphed here was the commercial motivation – so many billions of dollars were put into the development of general-purpose CPUs because of the substantial profits returned that specialized machines with extremely limited investment were never able to compete. The scientific conjecture that specialized machines would be faster on reasoning tasks than emulation never had a chance to be vindicated.

In the meantime, another kind of specialized processor has become successful. These are graphics processing units (GPUs) that are processors specialized for maniputating computer graphics and image processing. GPUs now have a significant share of the processor market and every laptop, for example, has both a CPU and a GPU. Furthermore, large parallel GPU machines are now widely used for deep learning applications. Scientifically, the concept of a logic processing unit (LPU) to do reasoning for AI and a GPU for graphics, image processing, and deep learning are both good ideas, but the latter has actually been realized in a dramatic way because of the commercial motivation, while the former has been in the dustbin of history for nearly 30 years.

So why did LPUs fail? The reason is that in the 1980s and up to now there has never been a commercial imperative to develop them. While logic has been influential at the academic level in AI research, few important implemented AI systems made much use of declarative languages, certainly not enough to encourage computer companies to commit the substantial development costs of developing LPUs (in contrast to GPUs that were motivated by the huge games industry). However, this situation should not continue indefinitely. AI applications are growing at an extraordinary rate and if the case that AI applications would greatly benefit from highly expressive Bach-like reasoning systems has merit, then the commercial opportunities for the development of LPUs will make the investment attractive and LPUs that are faster than emulation on CPUs will appear. This would lead to a future where computers commonly each contain CPUs, GPUs, LPUs, and, eventually, QPUs (quantum processing units). But, first, efficient and comprehensive implementations of Bach-like languages on conventional computers are needed and these languages applied in many applications to test the claim that such languages do greatly benefit AI.

Exercises

5.1 For the definitions given in Example 5.3.1, show the computations for the following.

(i) $(\mathit{permute}\ (1 \sharp 2 \sharp [], x))$.

(ii) $(\mathit{sorted}\ 2 \sharp 1 \sharp 3 \sharp [])$.

5.2 Let $P : \alpha \to \mathit{Bool}$ and $Q : \alpha \to \mathit{Bool}$ be two predicate constants. Give a proof of

$$\forall x.(Q\ x) \longrightarrow \exists x.\forall y.\neg((P\ y) \wedge \neg((P\ x) \wedge (Q\ x))).$$

5.3 Suppose that a belief theory includes

$$\bullet \boldsymbol{B}\varphi_1,\ \bullet^2 \boldsymbol{B}\varphi_2,\ \bullet^3 \boldsymbol{B}\varphi_3,\ \bullet^4 \boldsymbol{B}\varphi_4,\ \text{and}\ \bullet^5 \boldsymbol{B}\varphi_5$$

as local assumptions. Using the global assumption

$$\bullet \boldsymbol{B}\boldsymbol{\varphi} \longrightarrow \boldsymbol{B}\bullet\boldsymbol{\varphi},$$

show that, for each $i \in \{1, \ldots, 5\}$, $\boldsymbol{B}\bullet^i \varphi_i$ is a theorem of the belief theory.

Appendix A

Probability

THIS appendix presents some relevant concepts from probability theory. The presentation is narrowly focussed on the material needed to support a theory of empirical beliefs. As well as various basic concepts, there is a detailed discussion about probability kernels that underlie the belief representation requirements of agents. Proofs are usually given only for results that are new or not easily accessible in the literature.

A.1 Measurable Spaces and Measurable Functions

Notation.
$\mathbb{C}$ is the set of complex numbers.
$\mathbb{R}$ is the set of real numbers.
$\mathbb{Q}$ is the set of rationals.
$\mathbb{Z}$ is the set of integers.
$\mathbb{N}_0$ is the set of natural numbers, that is, non-negative integers.
$\mathbb{N}$ is the set of positive integers.
$\mathbb{B}$ is the set $\{\mathsf{T}, \mathsf{F}\}$ of booleans; T is true and F is false.

Let X and Y be sets. A *function* $f : X \to Y$ is a subset f of $X \times Y$ such that, for each $x \in X$, there is a unique $y \in Y$ such that $(x, y) \in f$, in which case one writes $y = f(x)$. (This definition identifies a function with its *graph* $\{(x, f(x)) \mid x \in X\}$.) The set X is called the *domain* and the set Y the *codomain* of the function. The *range* of the function is $f(X)$. Its *signature* is $X \to Y$.

Notation. Let I and X be sets and $f : I \to X$ a function. Sometimes the intention is to focus more on the range $f(I)$ of the function rather than the function f itself. In such cases, different terminology and notation is employed. A value $f(i)$ is instead denoted by x_i, and f is instead denoted by $(x_i)_{i \in I}$ and called an *indexed family* or, simply, a *family* of elements of X. Each $i \in I$ is called an *index* and the set I is called an *index set*. It may be true, and is often the case, that $x_i = x_j$, for some i and j such that $i \neq j$.

Thus, technically, an indexed family is just a function but denoted in such a way as to emphasize its range. The notation $(x_i)_{i \in I}$ should be carefully distinguished from $\{x_i\}_{i \in I}$ or, equivalently, $\{x_i \mid i \in I\}$ that denotes the set $f(I)$. If $I = \mathbb{N}$ (or $\mathbb{N}_0$), then $(x_i)_{i \in I}$ can

be interpreted as an (infinite) sequence of elements of X. If I is finite and ordered, then $(x_i)_{i\in I}$ can be interpreted as a (finite) sequence of elements of X.

Let $(X_i)_{i\in I}$ be an indexed family of subsets of a set X. Then $(X_i)_{i\in I}$ is a *partition* of X if $\bigcup_{i\in I} X_i = X$, $X_i \neq \emptyset$, for all $i \in I$, and $i \neq j$ implies $X_i \cap X_j = \emptyset$, for all $i, j \in I$.

It will be convenient to have λ-notation available as an (informal) mathematical notation.

Notation. Let X and Y be sets, x a variable ranging over values in X, and φ be an expression taking values in Y. The expression φ may or may not contain x as a free variable. Then the notation $\lambda x.\varphi$ denotes the function from X to Y defined by $x \mapsto \varphi$, for all $x \in X$. The understanding is that, for each element e of X, the corresponding value of the function is obtained by replacing each free occurrence of x in φ by e.

More generally, for $i = 1, \ldots, n$, let X_i be a set and x_i a variable ranging over values in X_i. The expression φ may or may not contain x_i as a free variable, for $i = 1, \ldots, n$. Then the notation $\lambda(x_1, \ldots, x_n).\varphi$ denotes the function from $X_1 \times \cdots \times X_n$ to Y defined by $(x_1, \ldots, x_n) \mapsto \varphi$, for all $x_1 \in X_1, \ldots, x_n \in X_n$. The understanding is that, for each element $(e_1, \ldots, e_n)$ of $X_1 \times \cdots \times X_n$, the corresponding value of the function is obtained by replacing each free occurrence of x_i in φ by e_i, for $i = 1, \ldots, n$.

The notation can be iterated. Thus let X, Y, and Z be sets, x a variable ranging over values in X, y a variable ranging over values in Y, and φ be an expression taking values in Z. The expression φ may or may not contain x or y as a free variable. Then the notation $\lambda x.\lambda y.\varphi$ denotes the function from X to Z^Y defined by $x \mapsto \lambda y.\varphi$, for all $x \in X$.

The λ-notation is also occasionally used in the infinite dimensional case. For example, $\lambda(x_1, x_2, \ldots).x_m : \prod_{n\in\mathbb{N}} X_n \to X_m$ is the canonical projection onto X_m.

The λ-notation is especially useful for compactly defining functions constructed in various ways from some other functions. As a typical example, see the function

$$\lambda(x_1, \ldots, x_n).\bigotimes_{i=1}^{n} \mu_i(x_i) : \prod_{i=1}^{n} X_i \to \mathcal{P}(\prod_{i=1}^{n} Y_i)$$

in the statement of Proposition A.2.17.

The *formal* treatment of λ-notation in higher-order logic is given, for the syntax, in Appendix B.1 and, for the semantics, in Appendix B.2.

The set of all subsets of a set X is denoted by 2^X. In effect, the power set of X has been identified with all predicates on X.

The important property of $\mathbb{B}$ is that it has two elements. These could just as easily be 0 (instead of F) and 1 (instead of T), or, alternatively, -1 (instead of F) and 1 (instead of T). In the following these three views of $\mathbb{B}$ are used interchangeably. For example, since the '2' in 2^X means the set $\{0, 1\}$, one can also write 2^X as $\mathbb{B}^X$.

Definition A.1.1. Let X be a set. A collection $\mathcal{A} \subseteq 2^X$ is called a *σ-algebra* on X if $X \in \mathcal{A}$, and $\mathcal{A}$ is closed under complementation and countable unions.

The *trivial σ-algebra* is $\{\emptyset, X\}$. At the opposite extreme, the power set 2^X of X is a σ-algebra of X.

Definition A.1.2. A *measurable space* is a pair $(X, \mathcal{A})$, where X is a set and $\mathcal{A}$ is a σ-algebra on X.

Each of $\mathbb{N}_0$, $\mathbb{N}$, and $\mathbb{B}$ is a measurable space with the σ-algebra being the respective set of all subsets.

Definition A.1.3. Let X be a set and $\mathcal{A} \subseteq 2^X$. Then the σ-algebra *generated by* $\mathcal{A}$ is the smallest σ-algebra containing $\mathcal{A}$ and is denoted by $\sigma(\mathcal{A})$.

The set of all subsets of a set X is a σ-algebra on X. Hence $\sigma(\mathcal{A})$ exists and is the intersection of all σ-algebras that contain $\mathcal{A}$.

Definition A.1.4. Let X be a set. A collection $\mathcal{A} \subseteq 2^X$ is called a *π-system* on X if $\mathcal{A}$ is closed under finite intersections.

Definition A.1.5. Let X be a set. A collection $\mathcal{A} \subseteq 2^X$ is called a *λ-system* on X if $X \in \mathcal{A}$; for all $A, B \in \mathcal{A}$ where $A \subseteq B$, $B \setminus A \in \mathcal{A}$; and $\mathcal{A}$ is closed under the union of a sequence of increasing sets.

Proposition A.1.1. *Let X be a set. A collection of subsets of X is a σ-algebra iff it is both a π-system and a λ-system.*

Proof. Straightforward. □

Proposition A.1.2. *(Monotone-class theorem) Let X be a set, $\mathcal{P}$ a π-system on X, and $\mathcal{L}$ a λ-system on X such that $\mathcal{P} \subseteq \mathcal{L}$. Then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

Proof. See [24, Theorem 1.8] or [83, Theorem 1.1]. □

Definition A.1.6. Let $(X, \mathcal{T})$ be a topological space. The *Borel σ-algebra* for $(X, \mathcal{T})$ is $\sigma(\mathcal{T})$. Each element of the Borel σ-algebra is called a *Borel set* (for X).

In other words, the Borel σ-algebra is the σ-algebra generated by all the open sets in the topology.

Notation. The Borel σ-algebra for a topological space $(X, \mathcal{T})$ is denoted by $\mathcal{B}(X)$; thus $\mathcal{B}(X) \triangleq \sigma(\mathcal{T})$.

Note that the symbol $\triangleq$ means 'stand(s) for'.

Definition A.1.7. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A function $f : X \to Y$ is *measurable* if $f^{-1}(B) \in \mathcal{A}$, for all $B \in \mathcal{B}$.

In Definition A.1.7, if $Y = \mathbb{R}^n$, then $\mathcal{B}$ is taken to be the Borel σ-algebra for $\mathbb{R}^n$.

Many σ-algebras are naturally generated by a function mapping into a measurable space.

Definition A.1.8. Let X be a set, $(Y, \mathcal{B})$ a measurable space, and $f : X \to Y$ a function. Then the σ-algebra *generated by* f is defined to be $\{f^{-1}(B) \mid B \in \mathcal{B}\}$, and is denoted by $\sigma(f)$.

Note that $\sigma(f)$ is the smallest σ-algebra on X for which f is measurable. More generally, a σ-algebra can be generated by an indexed family of functions.

Definition A.1.9. Let X be a set and I an index set. For all $i \in I$, let $(Y_i, \mathcal{B}_i)$ be a measurable space and $f_i : X \to Y_i$ a function. Then the σ-algebra *generated by* $(f_i)_{i\in I}$ is defined to be $\sigma(\bigcup_{i\in I} \sigma(f_i))$, and is denoted by $\sigma((f_i)_{i\in I})$.

Note that $\sigma((f_i)_{i\in I})$ is the smallest σ-algebra on X for which each f_i is measurable.

Here are two useful results that depend upon topological conditions.

Proposition A.1.3. *Let $(X, \mathcal{A})$ be a measurable space and $(Y, \mathcal{B})$ a measurable space, where Y is a metric space and $\mathcal{B}$ is its Borel σ-algebra. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of measurable functions from X to Y, $f : X \to Y$ a function, and suppose that $f_n(x) \to f(x)$, for all $x \in X$. Then f is measurable.*

Proof. See [83, Lemma 1.10] or [43, Theorem 4.2.2]. □

Proposition A.1.4. *Let $(X, \mathcal{A})$ be a measurable space, $(Y, \mathcal{B})$ a measurable space, where Y is a separable metric space and $\mathcal{B}$ its Borel σ-algebra, and $f : X \to Y$ a measurable function. Then $\{(x, f(x)) \mid x \in X\}$ is a measurable subset of $X \times Y$.*

Proof. Since Y is a separable metric space, for all $m \in \mathbb{N}$, there is a partition $\{A_n^{(m)}\}_{n\in\mathbb{N}}$ of Y by measurable sets of diameter less that $1/m$. Then

$$\{(x, f(x)) \mid x \in X\} = \bigcap_{m\in\mathbb{N}} \bigcup_{n\in\mathbb{N}} (f^{-1}(A_n^{(m)}) \times A_n^{(m)}) \in \mathcal{A} \otimes \mathcal{B},$$

since f is measurable. □

In other words, under some restrictions on the codomain, the graph of a measurable function is a measurable set.

Various ways of constructing measurable spaces from other measurable spaces are now considered.

Definition A.1.10. Let $(X, \mathcal{A})$ be a measurable space and $Y \subseteq X$. Then the *restriction* $\mathcal{A}|_Y$ of the σ-algebra of $\mathcal{A}$ by Y is defined by $\mathcal{A}|_Y = \{A \cap Y \mid A \in \mathcal{A}\}$.

Clearly, $\mathcal{A}|_Y$ is a σ-algebra on Y and so $(Y, \mathcal{A}|_Y)$ is a measurable space.

Here is the usual definition of composition of functions. (Later, in Appendix B.1, a reverse form of this will be introduced.)

Definition A.1.11. The composition function

$$\circ : (Y \to Z) \to (X \to Y) \to (X \to Z)$$

is defined by $(g \circ f)(x) = g(f(x))$, for all functions $f : X \to Y$ and $g : Y \to Z$, and $x \in X$.

Recall that the product $\prod_{i\in I} X_i$ of an indexed family $(X_i)_{i\in I}$ of sets is the set of all functions $f : I \to \bigcup_{i\in I} X_i$ such that $f(i) \in X_i$, for all $i \in I$. In other words, $\prod_{i\in I} X_i$ is the set of indexed families of the form $(x_i)_{i\in I}$, where $x_i \in X_i$, for all $i \in I$. If $I = \emptyset$, then $\prod_{i\in I} X_i = \{\emptyset\}$. (Note that $\emptyset$ *is* a function, usually denoted by (), in this context.) If X is a set and $n \in \mathbb{N}$, then X^n means $\prod_{i\in\{1,\ldots,n\}} X_i$, where $X_i = X$, for all $i = 1, \ldots, n$.

Definition A.1.12. Let I be an index set, $(X_i, \mathcal{A}_i)$ a measurable space, for all $i \in I$, $\prod_{i\in I} X_i$ the product of the X_i, and, for all $i \in I$, $\pi_i : \prod_{i\in I} X_i \to X_i$ the canonical projection defined by $\pi_i((x_i)_{i\in I}) = x_i$, for all $(x_i)_{i\in I} \in \prod_{i\in I} X_i$. The *product σ-algebra* $\bigotimes_{i\in I} \mathcal{A}_i$ on $\prod_{i\in I} X_i$ is defined to be $\sigma((\pi_i)_{i\in I})$, the σ-algebra generated by the set of projections. The measurable space $(\prod_{i\in I} X_i, \bigotimes_{i\in I} \mathcal{A}_i)$ is called a *product space*.

Let $(X, \mathcal{A})$ be a measurable space, I an index set, $(X_i, \mathcal{A}_i)$ a measurable space, for all $i \in I$, and $f : X \to \prod_{i\in I} X_i$ a function. Then it is clear that f is measurable iff $\pi_i \circ f : X \to X_i$ is measurable, for all $i \in I$.

Let I be an index set, $(X, \mathcal{A})$ and $(Y_i, \mathcal{B}_i)$, for all $i \in I$, measurable spaces, and $f_i : X \to Y_i$ a measurable function, for all $i \in I$. Define $\prod_{i\in I} f_i : X \to \prod_{i\in I} Y_i$ by $(\prod_{i\in I} f_i)(x) = (f_i(x))_{i\in I}$. Then $\prod_{i\in I} f_i$ is a measurable function. The notation $(f_i)_{i\in I}$ is more convenient than $\prod_{i\in I} f_i$ and generally is used instead.

Definition A.1.13. Let I be an index set, $(X_i, \mathcal{A}_i)$ a measurable space, for all $i \in I$, and $\coprod_{i\in I} X_i$ the sum (that is, disjoint union) of the X_i. The *sum σ-algebra* $\bigoplus_{i\in I} \mathcal{A}_i$ on $\coprod_{i\in I} X_i$ is defined to be $\{\coprod_{i\in I} A_i \mid A_i \in \mathcal{A}_i, \text{ for all } i \in I\}$. The measurable space $(\coprod_{i\in I} X_i, \bigoplus_{i\in I} \mathcal{A}_i)$ is called a *sum space*.

In the category of measurable spaces, sum is the coproduct.

Example A.1.1. Let $(X, \mathcal{A})$ be a measurable space and $\{*\}$ a singleton set. Given the σ-algebra $\{\{\}, \{*\}\}$, the set $\{*\}$ becomes a measurable space. Then the sum $X \sqcup \{*\}$ is a measurable space, called the *one point extension* of X. Explicitly, the σ-algebra on $X \sqcup \{*\}$ is $\mathcal{A} \cup \{A \sqcup \{*\} \mid A \in \mathcal{A}\}$.

Let $(X, \mathcal{A})$ be a measurable space, I an index set, $(X_i, \mathcal{A}_i)$ a measurable space, for all $i \in I$, and $f : \coprod_{i\in I} X_i \to X$ a function. Then it is clear that f is measurable iff $f|_{X_i} : X_i \to X$ is measurable, for all $i \in I$.

Let I be an index set, $(X_i, \mathcal{A}_i)$, for all $i \in I$, and $(Y, \mathcal{B})$ measurable spaces, and $f_i : X_i \to Y$ a measurable function, for $i \in I$. Define $\coprod_{i\in I} f_i : \coprod_{i\in I} X_i \to Y$ by $(\coprod_{i\in I} f_i)(x) = f_i(x)$, whenever $x \in X_i$. Then $\coprod_{i\in I} f_i$ is a measurable function.

Recall that the function space Y^X is the set of functions from X to Y, where X and Y are sets. The function space Y^X is equal to the product space $\prod_{x\in X} Y_x$, where $Y_x = Y$, for all $x \in X$ (that is, the product of $|X|$ copies of Y). In the following, 'product space' will be taken to be the primary concept and 'function space' a special kind of product space for which all the factors are identical.

Definition A.1.14. Let $(Y, \mathcal{B})$ be a measurable space and X a set. For each $x \in X$, the *evaluation map* $\pi_x : Y^X \to Y$ is defined by $\pi_x(f) = f(x)$, for all $f \in Y^X$.

The evaluation maps define the σ-algebra $\sigma((\pi_x)_{x\in X})$ on Y^X, the smallest σ-algebra for which each evaluation map is measurable. Thus $(Y^X, \sigma((\pi_x)_{x\in X}))$ is the measurable space of all functions from X to Y. If $U \subseteq Y^X$, then $\sigma((\pi_x)_{x\in X})|_U = \sigma((\pi_x|_U)_{x\in X})$.

Let $(X, \mathcal{A})$ be a measurable space. The function space $X^{\mathbb{N}}$ can be identified with the set of all infinite sequences whose elements belong to X. Similarly, the function space $X^{\mathbb{N}_0}$ can also be identified with the set of all infinite sequences whose elements belong to X, but with the indexing starting from 0.

Proposition A.1.5. *Let $(Y, \mathcal{B})$ and $(Z, \mathcal{C})$ be measurable spaces, X a set, and U a non-empty subset of Y^X. Consider U as a measurable space with the σ-algebra $\sigma((\pi_x)_{x\in X})|_U$. Then a function $g : Z \to U$ is measurable iff $\pi_x \circ g : Z \to Y$ is measurable, for all $x \in X$.*

Proof. See [83, Lemma 3.1]. □

Note that $\pi_x \circ g = \lambda z.g(z)(x)$.

A special case of the space of functions above is the set of all subsets of a set X, which is isomorphic to the function space $\mathbb{B}^X$. For this case, the canonical σ-algebra on $\mathbb{B}^X$ is the smallest σ-algebra generated by the evaluation maps $\pi_x : \mathbb{B}^X \to \mathbb{B}$ defined by $\pi_x(s) = s(x)$, for all $s \in \mathbb{B}^X$. (If s is interpreted as a subset of X, this can be written as $\pi_x(s) = \mathsf{T}$, if $x \in s$, and $\pi_x(s) = \mathsf{F}$, if $x \notin s$.)

Of more interest for practical applications is the set $\mathcal{F}_X$ of all finite subsets of X. Thus $\mathcal{F}_X$ is isomorphic to $\{s \in \mathbb{B}^X \mid s(x) = \mathsf{F}, \text{ for all but finitely many } x \in X\}$. In fact, applications typically involve the set of subsets of some finite subset F of X; that is,

$$\{s \in \mathbb{B}^X \mid s(x) = \mathsf{F}, \text{ for all } x \in X \setminus F\}.$$

Note the natural bijection between this set and $\mathbb{B}^F$.

Let X be a set and $\sim$ an equivalence relation on X. For all $x \in X$, let $[x] \triangleq \{y \in X \mid y \sim x\}$. Each $[x]$ is called an equivalence class. Let $X/\sim \triangleq \{[x] \mid x \in X\}$ be the set of all equivalence classes. $X/\sim$ is called a quotient space. The function $p : X \to X/\sim$ defined by $p(x) = [x]$, for all $x \in X$, is a surjection. Conversely, for any surjection $p : X \to Y$, there is an equivalence relation $\sim$ on X defined by $x \sim y$ if $p(x) = p(y)$, for all $x, y \in X$. Furthermore, Y is isomorphic to $X/\sim$. In this case, the quotient space $X/\sim$ is also denoted by X/p.

Now let $(X, \mathcal{A})$ be a measurable space and $\sim$ an equivalence relation on X. Let $\mathcal{A}/\sim \triangleq \{A \subseteq X/\sim \mid \bigcup_{[x]\in A}[x] \in \mathcal{A}\}$. Clearly $\mathcal{A}/\sim$ is a σ-algebra on $X/\sim$. Thus $(X/\sim, \mathcal{A}/\sim)$ is a measurable space.

Definition A.1.15. Let $(X, \mathcal{A})$ be a measurable space and $\sim$ an equivalence relation on X. Then the *quotient space* is $(X/\sim, \mathcal{A}/\sim)$.

The next four propositions are relevant to the construction of various quotients discussed later.

Proposition A.1.6. *Let $(X, \mathcal{A})$ be a measurable space and $f : \coprod_{n\in\mathbb{N}_0} X^n \to \mathbb{B}^X$ be defined by*

$$f(x_1, \ldots, x_n) = \{x \mid x = x_i, \text{ for some } i \in \{1, \ldots, n\}\},$$

for all $n \in \mathbb{N}_0$ and $(x_1, \ldots, x_n) \in X^n$. Then f is measurable.

Proof. Informally, if $\coprod_{n\in\mathbb{N}_0} X^n$ is regarded as the sets of all (finite) lists of elements of X, then f maps a list to the set of elements in the list.

To show f is measurable, it suffices to show that

$$f|_{X^n} : X^n \to \mathbb{B}^X$$

is measurable, for all $n \in \mathbb{N}_0$. For this, by Proposition A.1.5, it suffices to show that

$$\pi_x \circ f|_{X^n} : X^n \to \mathbb{B}$$

is measurable, for all $n \in \mathbb{N}_0$ and $x \in X$. Now

$$\begin{aligned}
&(\pi_x \circ f|_{X^n})^{-1}(\{\mathsf{F}\}) \\
&= \{(x_1, \ldots, x_n) \in X^n \mid x \neq x_i, \text{ for all } i = 1, \ldots, n\} \\
&= (X \setminus \{x\})^n,
\end{aligned}$$

which is a measurable subset of X^n (assuming singleton subsets of X are measurable). It follows that $\pi_x \circ f|_{X^n}$ is measurable, for all $n \in \mathbb{N}_0$ and $x \in X$, and hence f is measurable. □

Proposition A.1.7. *Let $(X, \mathcal{A})$ be a measurable space and $f : \coprod_{n \in \mathbb{N}_0} X^n \to \mathbb{N}_0^X$ be defined by*

$$f(x_1, \ldots, x_n) = \lambda x.|\{i \in \{1, \ldots, n\} \mid x = x_i\}|,$$

for all $n \in \mathbb{N}_0$ and $(x_1, \ldots, x_n) \in X^n$. Then f is measurable.

Proof. Informally, f maps a list to the multiset obtained from the list by ignoring the order of occurrence of elements.

To show f is measurable, it suffices to show that

$$f|_{X^n} : X^n \to \mathbb{N}_0^X$$

is measurable, for all $n \in \mathbb{N}_0$. For this, by Proposition A.1.5, it suffices to show that

$$\pi_x \circ f|_{X^n} : X^n \to \mathbb{N}_0$$

is measurable, for all $n \in \mathbb{N}_0$ and $x \in X$. Let $k \in \mathbb{N}_0$. Then

$$\begin{aligned}
&(\pi_x \circ f|_{X^n})^{-1}(\{k\}) \\
&= \{(x_1, \ldots, x_n) \in X^n \mid |\{i \in \{1, \ldots, n\} \mid x = x_i\}| = k\} \\
&= \bigcup_{s \subseteq \{1, \ldots, n\} \,:\, |s| = k} \left(\prod_{i \in s} \{x\} \times \prod_{i \notin s} (X \setminus \{x\})\right)
\end{aligned}$$

which is a measurable subset of X^n (assuming singleton subsets of X are measurable). It follows that $\pi_x \circ f|_{X^n}$ is measurable, for all $n \in \mathbb{N}_0$ and $x \in X$, and hence f is measurable. □

Proposition A.1.8. *Let $(X, \mathcal{A})$ be a measurable space and $f : \mathbb{N}_0^X \to \mathbb{B}^X$ be defined by*

$$f(m)(x) = \begin{cases} \mathsf{F} & \text{if } m(x) = 0 \\ \mathsf{T} & \text{otherwise,} \end{cases}$$

for all $m \in \mathbb{N}_0^X$ and $x \in X$. Then f is measurable.

Proof. Informally, f maps a multiset to the set obtained from the multiset by ignoring the multiplicity of elements.

Let $\pi_x : \mathbb{B}^X \to \mathbb{B}$ and $\pi'_x : \mathbb{N}_0^X \to \mathbb{N}_0$ be the respective evaluation maps, for all $x \in X$. It suffices to show that $\pi_x \circ f : \mathbb{N}_0^X \to \mathbb{B}$ is measurable, for all $x \in X$. But this is obvious since $(\pi_x \circ f)^{-1}(\{\mathsf{F}\}) = (\pi'_x)^{-1}(\{0\})$ is measurable, for all $x \in X$. □

Proposition A.1.9. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, where X is countable, and $f : X \to Y$ a measurable function. Define $\overline{f} : \mathbb{B}^X \to \mathbb{B}^Y$ by*

$$\overline{f}(s)(y) = \begin{cases} \mathsf{T} & \text{if } y = f(x) \text{ and } s(x) = \mathsf{T}, \text{ for some } x \in X \\ \mathsf{F} & \text{otherwise,} \end{cases}$$

for all $s \in \mathbb{B}^X$ and $y \in Y$. Then $\overline{f}$ is measurable.

Proof. It suffices to show that $\pi_y \circ \overline{f} : \mathbb{B}^X \to \mathbb{B}$ is measurable, for all $y \in Y$. Now

$$\begin{aligned} & (\pi_y \circ \overline{f})^{-1}(\{\mathsf{T}\}) \\ &= \overline{f}^{-1}(\pi_y^{-1}(\{\mathsf{T}\})) \\ &= \overline{f}^{-1}(\{t \in \mathbb{B}^Y \mid t(y) = \mathsf{T}\} \\ &= \bigcup_{x \in f^{-1}(\{y\})} \{s \in \mathbb{B}^X \mid s(x) = \mathsf{T}\}, \end{aligned}$$

which is a countable union of measurable sets. It follows that $\overline{f}$ is measurable. □

A.2 Probability Measures and Probability Kernels

Definition A.2.1. Let $(X, \mathcal{A})$ be a measurable space. A function $\mu : \mathcal{A} \to [0, \infty]$ is a *measure* (on $(X, \mathcal{A})$) if it is countably additive (that is, if $(B_i)_{i=1}^{\infty}$ is a sequence of pairwise disjoint sets in $\mathcal{A}$, then $\mu(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mu(B_i)$).

A measure μ is a *probability measure* if $\mu(X) = 1$.

To avoid trivial measures, it is assumed that there exists $A \in \mathcal{A}$ such that $\mu(A) < \infty$, from which it is easy to prove that $\mu(\emptyset) = 0$.

Next comes the definition of a probability space.

Definition A.2.2. A *measure space* is a triple $(X, \mathcal{A}, \mu)$, where $(X, \mathcal{A})$ is a measurable space and μ is a measure on $(X, \mathcal{A})$.

A *probability space* is a triple $(X, \mathcal{A}, \mu)$, where $(X, \mathcal{A})$ is a measurable space and μ is a probability measure on $(X, \mathcal{A})$.

If $(X, \mathcal{A}, \mu)$ is a measure space, a statement about $x \in X$ is said to hold *μ-almost everywhere*, or μ-a.e., if it holds for all $x \notin A$, for some $A \in \mathcal{A}$ with $\mu(A) = 0$. If the relevant measure is clear from the context, it is denoted simply by a.e. The phrase 'μ-almost all $x \in X$' is also used in the same context.

Example A.2.1. If $(X, \mathcal{A})$ is a measurable space, then the *Dirac measure* δ_x (*at* x), for some $x \in X$, is defined by $\delta_x(A) = \mathbf{1}_A(x)$, for all $A \in \mathcal{A}$. (Here $\mathbf{1}_A$ is the indicator (that is, characteristic) function for A.) Thus

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise,} \end{cases}$$

for all $A \in \mathcal{A}$. The set of all Dirac measures on X is denoted by $\Delta(X)$, that is, $\Delta(X) \triangleq \{\delta_x \mid x \in X\}$. A Dirac measure is a probability measure.

Example A.2.2. Let I be a countable index set, $(a_i)_{i \in I}$ an indexed family of non-negative real numbers such that $\sum_{i \in I} a_i = 1$, and μ_i a probability measure on $(X, \mathcal{A})$, for $i \in I$. Then $\sum_{i \in I} a_i \mu_i$ is a probability measure on $(X, \mathcal{A})$ called a *mixture measure*. In the literature, a mixture measure is often called a *mixture model*.

In addition, let $(x_i)_{i \in I}$ be an indexed family of elements of X. Then the probability measure $\sum_{i \in I} a_i \delta_{x_i}$ is called a *Dirac mixture measure*.

Typically, in applications, I is finite.

Example A.2.3. Let X be a set. *Counting measure* c is defined on the σ-algebra of all subsets of X by

$$c(A) = \begin{cases} \text{cardinality of } A & \text{if } A \text{ is finite} \\ \infty & \text{otherwise,} \end{cases}$$

for all $A \subseteq X$.

Example A.2.4. Let $X \triangleq \{x_1, \ldots, x_n\}$ be a finite set and $(a_i)_{i=1}^m$ a sequence of non-negative real numbers such that $\sum_{i=1}^n a_i = 1$. (The x_i are pairwise distinct, but the a_i may not be.) Then the *categorical measure* μ (determined by $(a_i)_{i=1}^n$) is defined on the σ-algebra 2^X by $\mu(\{x_i\}) = a_i$, for $i = 1, \ldots, n$. Hence $\mu(A) = \sum_{x_i \in A} a_i$, for all $A \subseteq X$. Equivalently, $\mu = \sum_{i=1}^n a_i \delta_{x_i}$. Each categorical measure is a probability measure. Conversely, every probability measure μ on a measurable space $(X, 2^X)$, where $X \triangleq \{x_1, \ldots, x_n\}$ is finite, is a categorical measure: define $a_i = \mu(\{x_i\})$, for $i = 1, \ldots, n$; then $\mu = \sum_{i=1}^n a_i \delta_{x_i}$.

Example A.2.5. On $\mathbb{R}^n$, the standard measure is *Lebesgue measure* [87] on the Borel σ-algebra for $\mathbb{R}^n$. Lebesgue measure is usually denoted by λ.

The collection of all probability measures on some measurable space will often be needed.

Notation. The set of all measures on a measurable space $(X, \mathcal{A})$ is denoted by $\mathcal{M}(X)$, with the $\mathcal{A}$ understood. Similarly, the set of all probability measures on $(X, \mathcal{A})$ is denoted by $\mathcal{P}(X)$.

$\mathcal{M}(X)$ becomes a measurable space when given the σ-algebra generated by the indexed family of functions $(\pi_A)_{A \in \mathcal{A}}$, where $\pi_A : \mathcal{M}(X) \to [0, \infty]$ is defined by $\pi_A(\mu) = \mu(A)$, for all $\mu \in \mathcal{M}(X)$. Note that $\mathcal{P}(X)$ is a measurable subset of $\mathcal{M}(X)$ (since $\mathcal{P}(X) = \pi_X^{-1}(\{1\})$ and π_X is measurable).

Definition A.2.3. A *simple* function is any linear combination of measurable indicator functions.

Thus a simple function has the form $\sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$, where $c_i \in \mathbb{R}$ and A_i is measurable, for $i = 1, \ldots, n$.

Proposition A.2.1. *Let $(X, \mathcal{A})$ be a measurable space and $f : X \to \mathbb{R}$ a non-negative measurable function. Then there exists an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of non-negative simple functions such that $\lim_{n\to\infty} f_n = f$.*

Proof. For all $n \in \mathbb{N}$, define $f_n : X \to \mathbb{R}$ by $f_n(x) = 2^{-n}\lfloor 2^n f(x) \rfloor \wedge n$, for all $x \in X$. □

Definition A.2.4. If $(X, \mathcal{A}, \mu)$ is a measure space and $f = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ a non-negative simple function, then the *integral* $\int_X f \, d\mu$ of f is defined by $\int_X f \, d\mu = \sum_{i=1}^{n} c_i \mu(A_i)$.

If $f : X \to \mathbb{R}$ is a non-negative measurable function, then the *integral* $\int_X f \, d\mu$ of f is defined by

$$\int_X f \, d\mu = \lim_{n\to\infty} \int_X f_n \, d\mu,$$

where $(f_n)_{n\in\mathbb{N}}$ is a sequence of non-negative simple functions such that $\lim_{n\to\infty} f_n = f$.

The existence of the sequence $(f_n)_{n\in\mathbb{N}}$ in the definition of the integral is given by Proposition A.2.1. It can be shown that this definition is independent of the choice of $(f_n)_{n\in\mathbb{N}}$.

Definition A.2.5. Let $(X, \mathcal{A}, \mu)$ be a measure space. A measurable function $f : X \to \mathbb{R}$ is *integrable* if $\int_X |f| \, d\mu < \infty$.

If f is integrable, then f can be written in the form $f = f^+ - f^-$, where f^+ and f^- are non-negative integrable functions.

Definition A.2.6. Let $(X, \mathcal{A}, \mu)$ be a measure space and $f : X \to \mathbb{R}$ an integrable function. Then the *integral* $\int_X f \, d\mu$ of f is defined by

$$\int_X f \, d\mu = \int_X f^+ \, d\mu - \int_X f^- \, d\mu.$$

Example A.2.6. Let $(X, 2^X, c)$ be a measure space, where X is countable and c is counting measure. Suppose that $f : X \to \mathbb{R}$. Any such function f is measurable. Then f is integrable (with respect to $(X, 2^X, c)$) if and only if $\sum_{x\in X} |f(x)| < \infty$, and then $\int_X f \, dc = \sum_{x\in X} f(x)$. Thus, for counting measure, integrals reduce to sums.

Proposition A.2.2. *(Monotone convergence theorem) Let $(X, \mathcal{A}, \mu)$ be a measure space, $f : X \to \mathbb{R}$ a non-negative measurable function, and $(f_n)_{n\in\mathbb{N}}$ an increasing sequence of non-negative measurable functions such that $\lim_{n\to\infty} f_n = f$. Then $\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu$.*

Proof. See [83, Theorem 1.19] or [87, Theorem 4.20]. □

Proposition A.2.3. *(Dominated convergence theorem) Let $(X, \mathcal{A}, \mu)$ be a measure space, $f : X \to \mathbb{R}$ a measurable function, $(f_n)_{n\in\mathbb{N}}$ a sequence of measurable functions such that $\lim_{n\to\infty} f_n = f$, and $g : X \to \mathbb{R}$ an integrable function such that $|f_n| \leq g$, for all $n \in \mathbb{N}$. Then f is integrable and $\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu$.*

Proof. See [83, Theorem 1.21] or [87, Corollary 6.26]. □

Marginal probability measures will be useful.

Definition A.2.7. Let $(X_i, \mathcal{A}_i)$ be a measurable space, for $i = 1, \ldots, n$, and $\mu : \mathcal{P}(\prod_{i=1}^n X_i)$ a probability measure. Then the *marginal probability measure* $\mu_i : \mathcal{P}(X_i)$ (*for μ with respect to X_i*) is defined by

$$\mu_i(A_i) = \mu(X_1 \times \cdots \times X_{i-1} \times A_i \times X_{i+1} \times \cdots \times X_n),$$

for all $A_i \in \mathcal{A}_i$ and for $i = 1, \ldots, n$.

It is easy to show that each μ_i is actually a probability measure. In fact, let $\pi_i : \prod_{j=1}^n X_j \to X_i$ be the canonical projection, for $i = i, \ldots, n$. Then $\mu_i = \mu \circ \pi_i^{-1}$, for $i = i, \ldots, n$.

Now the discussion turns to kernels.

Definition A.2.8. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A function $\mu : X \to \mathcal{M}(Y)$ is called a *kernel* if μ is measurable.

If $\mu : X \to \mathcal{P}(Y)$, then μ is a *probability kernel*.

In the literature, a probability kernel is also known as a *stochastic kernel*, a *Markov kernel*, or a *transition kernel* [87, Definition 8.25], [83, p.20], [24, p.39].

The concept of a probability kernel is the most central belief representation concept in this book: agents, environments, schemas, transition models, and observation models are all (sequences of) probability kernels; empirical beliefs are probability kernels.

Note. It will be convenient sometimes to regard a probability measure as a special case of a probability kernel for which X is a distinguished singleton set $\{*\}$ with the σ-algebra $\{\{\}, \{*\}\}$ (in which case the set of probability kernels from X to $\mathcal{P}(Y)$ can be identified with $\mathcal{P}(Y)$).

Proposition A.2.4. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and $\mu : X \to \mathcal{P}(Y)$ a function. Then μ is a probability kernel iff, for all $B \in \mathcal{B}$, the function*

$$\lambda x.\mu(x)(B) : X \to \mathbb{R}$$

is measurable.

Proof. This follows directly from the definition of the σ-algebra on $\mathcal{P}(Y)$ and Proposition A.1.5. □

Proposition A.2.5. *Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$ and $(Z, \mathcal{C})$ be measurable spaces, and $\mu : X \times Y \to \mathcal{P}(Z)$ a probability kernel. Then, for all $x \in X$, $\lambda y.\mu(x, y) : Y \to \mathcal{P}(Z)$ is a probability kernel.*

Proof. By Proposition A.2.4, it suffices to show that, for all $x \in X$ and $C \in \mathcal{C}$, $\lambda y.\mu(x,y)(C) : Y \to \mathbb{R}$ is measurable. But this is obvious since μ is a probability kernel and so, for all $C \in \mathcal{C}$, $\lambda(x,y).\mu(x,y)(C) : X \times Y \to \mathbb{R}$ is measurable. □

The following property will be useful in later proofs.

Proposition A.2.6. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $\mu : X \to \mathcal{P}(Y)$ a probability kernel, and $f : X \times Y \to [0, \infty]$ a measurable function. Then $\lambda x. \int_Y \lambda y. f(x,y)\, d\mu(x) : X \to [0, \infty]$ is measurable.*

Proof. See [87, Lemma 14.20]. □

Marginal probability kernels can also be defined.

Definition A.2.9. Let $(X_i, \mathcal{A}_i)$ be a measurable space, for $i = 0, \ldots, n$, $\mu : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$ a probability kernel, and $\pi_i : \prod_{j=1}^n X_j \to X_i$ the canonical projection, for $i = i, \ldots, n$. Then the *marginal probability kernel* $\mu_i : X_0 \to \mathcal{P}(X_i)$ (*for μ with respect to X_i*) is defined by

$$\mu_i = \lambda x_0.(\mu(x_0) \circ \pi_i^{-1}),$$

for $i = 1, \ldots, n$.

Since each π_i is measurable, Proposition A.10.1 shows that μ_i is a probability kernel.

Much use will be made of piecewise-constant functions, since that is a convenient form for many probability kernels. A function $f : X \to Y$ is *piecewise-constant* if $|f(X)|$ is finite. Equivalently, a function $f : X \to Y$ is piecewise-constant iff there exists a partition $(X_i)_{i=1}^n$ of X and a subset $\{y_1, \ldots, y_n\}$ of Y such that $f(x) = y_i$, for all $x \in X_i$ and $i = 1, \ldots, n$. Thus the definition of a piecewise-constant function f has the form:

$\forall x.(f(x) =$
  if $x \in X_1$ *then* y_1
  else if $x \in X_2$ *then* y_2
    $\vdots$
  else if $x \in X_{n-1}$ *then* y_{n-1}
  else y_n).

Now, for all $i = 1, \ldots, n-1$, let $p_i : X \to \mathbb{B}$ be any predicate having the property that $p_i(x) = \mathsf{T}$ iff $x \in X_i$, for all $x \in X \setminus \bigcup_{j=1}^{i-1} X_j$. Then the definition of f can be written in the equivalent form:

$\forall x.(f(x) =$
  if $p_1(x)$ *then* y_1
  else if $p_2(x)$ *then* y_2
    $\vdots$
  else if $p_n(x)$ *then* y_{n-1}
  else y_n).

Note that this last form can also be written more compactly as

$$f = \lambda x.\mathit{if}\ p_1(x)\ \mathit{then}\ y_1\ \mathit{else\ if}\ p_2(x)\ \mathit{then}\ y_2\ \ \ldots\ \ \mathit{else\ if}\ p_{n-1}(x)\ \mathit{then}\ y_{n-1}\ \mathit{else}\ y_n.$$

This is the form in which piecewise-constant functions will be considered throughout this book. Typically, each p_i belongs to some fixed class of predicates and each function of the above form can be considered to be a potential approximation to a function in a (larger) class of functions from X to Y.

Example A.2.7. Here are some kinds of probability kernels that will be prominent in applications.

A probability kernel $\mu : X \to \mathcal{P}(Y)$ is called a *Dirac probability kernel* if $\mu(X) \subseteq \Delta(Y)$.

A probability kernel $\mu : X \to \mathcal{P}(Y)$ is called *piecewise-constant* if it is a piecewise-constant function. A common form for the definition of a piecewise-constant probability kernel is

$$\mu = \lambda x.\mathit{if}\ p_1(x)\ \mathit{then}\ \nu_1\ \mathit{else\ if}\ p_2(x)\ \mathit{then}\ \nu_2\ \ \ldots\ \ \mathit{else\ if}\ p_{n-1}(x)\ \mathit{then}\ \nu_{n-1}\ \mathit{else}\ \nu_n.$$

Here, $p_i : X \to \mathbb{B}$, for $i = 1, \ldots, n-1$, and $\nu_i \in \mathcal{P}(Y)$, for $i = 1, \ldots, n$. This form of definition is called a *decision list* and is particularly useful when X is structured.

Use will also be made of the 'fusion' of probability kernels.

Definition A.2.10. Let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$, and $(X_2, \mathcal{A}_2)$ be measurable spaces, and $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels. The *fusion* $\mu_1 \odot \mu_2 : X_0 \to \mathcal{P}(X_2)$ of μ_1 and μ_2 is defined by

$$(\mu_1 \odot \mu_2)(x_0) = \lambda A_2. \int_{X_1} \lambda x_1.\mu_2(x_0, x_1)(A_2)\ d\mu_1(x_0),$$

for all $x_0 \in X_0$.

In the literature ([83, p.21], [87, Definition 14.25]), fusion of probability kernels is usually called composition. The new terminology of 'fusion' has been introduced here to avoid confusion with ordinary composition of functions.

Fusion of probability kernels is well-defined since $\lambda x_1.\mu_2(x_0, x_1)(A_2) : X_1 \to \mathbb{R}$ is integrable, for all $x_0 \in X_0$ and $A_2 \in \mathcal{A}_2$, and clearly $(\mu_1 \odot \mu_2)(x_0) \in \mathcal{P}(X_2)$, for all $x_0 \in X_0$. Note that, for all $x_0 \in X_0$ and $A_2 \in \mathcal{A}_2$, $(\mu_1 \odot \mu_2)(x_0)(A_2)$ is the expected value of $\lambda x_1.\mu_2(x_0, x_1)(A_2)$ with respect to the probability measure $\mu_1(x_0)$.

Of course, in the preceding definition, A_2 ranges over all elements (that is, measurable sets) in $\mathcal{A}_2$. In Proposition A.2.8 below, it is shown that the fusion of probability kernels is again a probability kernel.

A special case of Definition A.2.10 will be useful. Suppose that X_0 is a singleton set. Then $\mu_1 : X_0 \to \mathcal{P}(X_1)$ can be identified in the obvious way with a probability measure also denoted by μ_1 and having signature $\mathcal{P}(X_1)$. Also $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ can be identified with a probability kernel also denoted by μ_2 and having signature $X_1 \to \mathcal{P}(X_2)$. In this case, the fusion $\mu_1 \odot \mu_2 : \mathcal{P}(X_2)$ of μ_1 and μ_2 is defined by

$$\mu_1 \odot \mu_2 = \lambda A_2. \int_{X_1} \lambda x_1.\mu_2(x_1)(A_2)\ d\mu_1.$$

Note that, for all $x_0 \in X_0$, $(\mu_1 \odot \mu_2)(x_0) = \mu_1(x_0) \odot \lambda x_1.\mu_2(x_0, x_1)$.

A good way to think about fusion is that it is a generalized mixture operation, as the following example illustrates.

Example A.2.8. Consider any categorical probability measure $\pi : \mathcal{P}(I)$, where I is a finite set with the σ-algebra 2^I. Let $\pi_i \triangleq \pi(\{i\})$, for $i \in I$. Then $\sum_{i \in I} \pi_i = 1$ and $\pi_i \geq 0$, for $i \in I$.

Consider also any measurable space $(X, \mathcal{A})$ and probability kernel $\mu : I \to \mathcal{P}(X)$. Then one can form the fusion of π and μ which is the probability measure $\pi \odot \mu : \mathcal{P}(X)$ defined by

$$(\pi \odot \mu)(A) = \int_I \lambda i.\mu(i)(A)\ d\pi = \sum_{i \in I} \pi_i \mu(i)(A),$$

for all $A \in \mathcal{A}$. Hence $\pi \odot \mu = \sum_{i \in I} \pi_i \mu(i)$. Note that $\sum_{i \in I} \pi_i \mu(i)$ is a mixture measure.

So fusion is just a generalization of the concept of mixing to the case where I and π are arbitrary.

Example A.2.9. Let $(A, \mathcal{A})$ be an action space and $(S, \mathcal{S})$ a state space. Consider (a component of) a transition model which is a probability kernel

$$\tau : A \times S \to \mathcal{P}(S).$$

Suppose $\gamma : \mathcal{P}(S)$ is the current probability measure on states. Then, for all $a \in A$, the next probability measure on states is $\gamma \odot \lambda s.\tau(a, s)$, where

$$\gamma \odot \lambda s.\tau(a, s) = \lambda T. \int_S \lambda s.\tau(a, s)(T)\ d\gamma.$$

As the next result shows, some transition models do not change the empirical belief.

Proposition A.2.7. *Let $(X_0, \mathcal{A}_0)$ and $(X_1, \mathcal{A}_1)$ be measurable spaces, and $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_1)$ probability kernels. Suppose that μ_2 is defined by $\mu_2(x_0, x_1) = \delta_{x_1}$, for all $x_0 \in X_0$ and $x_1 \in X_1$. Then $\mu_1 \odot \mu_2 = \mu_1$.*

Proof. For all $x_0 \in X_0$ and $A_1 \in \mathcal{A}_1$,

$$\begin{aligned}
&(\mu_1 \odot \mu_2)(x_0)(A_1) \\
&= \int_{X_1} \lambda x_1.\mu_2(x_0, x_1)(A_1)\ d\mu_1(x_0) \\
&= \int_{X_1} \lambda x_1.\delta_{x_1}(A_1)\ d\mu_1(x_0) \\
&= \int_{X_1} \lambda x_1.\mathbf{1}_{A_1}(x_1)\ d\mu_1(x_0) \\
&= \mu_1(x_0)(A_1).
\end{aligned}$$

□

Extensions of Definition A.2.10 to contexts where some of the domain arguments are missing will also be useful. For example, let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$, $(X_2, \mathcal{A}_2)$, $(X_3, \mathcal{A}_3)$, and $(X_4, \mathcal{A}_4)$ be measurable spaces, and $\mu_1 : X_0 \times X_1 \to \mathcal{P}(X_2)$ and $\mu_2 : X_0 \times X_3 \times X_1 \times X_2 \to \mathcal{P}(X_4)$ probability kernels. (So domain argument X_3 is missing from μ_1.) Consider the probability kernel $\lambda(x_0, x_3, x_1).\mu_1(x_0, x_1) : X_0 \times X_3 \times X_1 \to \mathcal{P}(X_2)$. Then $\mu_1 \odot \mu_2 : X_0 \times X_3 \times X_1 \to \mathcal{P}(X_4)$ is defined to be $(\lambda(x_0, x_3, x_1).\mu_1(x_0, x_1)) \odot \mu_2$. Thus

$$(\mu_1 \odot \mu_2)(x_0, x_3, x_1) = \lambda A_4. \int_{X_2} \lambda x_2.\mu_2(x_0, x_3, x_1, x_2)(A_4)\, d\mu_1(x_0, x_1),$$

for all $x_0 \in X_0$, $x_3 \in X_3$, and $x_1 \in X_1$.

The fusion of probability kernels results in a probability kernel.

Proposition A.2.8. *Let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$, and $(X_2, \mathcal{A}_2)$ be measurable spaces, and $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels. Then $\mu_1 \odot \mu_2 : X_0 \to \mathcal{P}(X_2)$ is a probability kernel.*

Proof. Clearly, by the monotone convergence theorem (Proposition A.2.2), $(\mu_1 \odot \mu_2)(x_0) \in \mathcal{P}(X_2)$, for all $x_0 \in X_0$. Thus $\mu_1 \odot \mu_2$ is well-defined.

For all $A_3 \in \mathcal{A}_3$, $\lambda x_0.(\mu_1 \odot \mu_2)(x_0)(A_2) = \lambda x_0. \int_{X_1} \lambda x_1.\mu_2(x_0, x_1)(A_2)\, d\mu_1(x_0)$ is measurable, by Proposition A.2.6. Hence, by Proposition A.2.4, $\mu_1 \odot \mu_2$ is a probability kernel. □

If $\mu : X \to \mathcal{P}(Y)$ is a probability kernel, then there is a natural associated probability kernel from $\mathcal{P}(X)$ to $\mathcal{P}(Y)$.

Proposition A.2.9. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and $\mu : X \to \mathcal{P}(Y)$ a probability kernel. Then $\lambda\gamma.(\gamma \odot \mu) : \mathcal{P}(X) \to \mathcal{P}(Y)$ is a probability kernel.*

Proof. It has to be shown that $\lambda\gamma.(\gamma \odot \mu)$ is measurable. For this, by Proposition A.2.4, it suffices to show that $\lambda\gamma.(\gamma \odot \mu)(B) : \mathcal{P}(X) \to \mathbb{R}$, that is, $\lambda\gamma. \int_X \lambda x.\mu(x)(B)\, d\gamma$, is measurable, for all $B \in \mathcal{B}$.

By Proposition A.2.1, for all $B \in \mathcal{B}$, there exists a sequence $(f_n)_{n\in\mathbb{N}}$ of simple functions such that $\lim_{n\to\infty} f_n = \lambda x.\mu(x)(B)$. Since the σ-algebra on $\mathcal{P}(X)$ is, by definition, the σ-algebra generated by the indexed family $(\lambda\gamma.\gamma(A) : \mathcal{P}(X) \to \mathbb{R})_{A\in\mathcal{A}}$, it follows that $\lambda\gamma. \int_X f_n\, d\gamma$ is measurable, for all $n \in \mathbb{N}$. Also the monotone convergence theorem (Proposition A.2.2) shows that, for all $B \in \mathcal{B}$ and $\gamma \in \mathcal{P}(X)$, $\lim_{n\to\infty} \int_X f_n\, d\gamma = \int_X \lambda x.\mu(x)(B)\, d\gamma$. Hence, by Proposition A.1.3, for all $B \in \mathcal{B}$, $\lambda\gamma. \int_X \lambda x.\mu(x)(B)\, d\gamma$, is measurable. □

Proposition A.2.10. *Let $(X, \mathcal{A})$ be a measurable space, and μ and ν bounded measures on $\mathcal{A}$. Suppose that $\mathcal{C}$ is a π-system on X such that $X \in \mathcal{C}$ and $\sigma(\mathcal{C}) = \mathcal{A}$. Then $\mu = \nu$ iff $\mu(A) = \nu(A)$, for all $A \in \mathcal{C}$.*

Proof. See [83, Lemma 1.17]. □

Definition A.2.11. Let $(X, \mathcal{A}, \mu)$ be a measure space. The measure μ is *σ-finite* if there exists a sequence $(X_i)_{i=1}^{\infty}$ of measurable subsets such that $\mu(X_i) < \infty$, for $i = 1, 2, \ldots$, and $X = \bigcup_{i=1}^{\infty} X_i$. In this case, the measure space $(X, \mathcal{A}, \mu)$ is also said to be *σ-finite*.

The measure μ is *finite* if $\mu(X)$ is finite. In this case, the measure space $(X, \mathcal{A}, \mu)$ is also said to be *finite*.

For the definition of a σ-finite measure, the sets in $(X_i)_{i=1}^{\infty}$ can, of course, be assumed to be pairwise disjoint, if desired.

Here is a uniqueness result that will be useful.

Proposition A.2.11. *Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ non-negative integrable functions such that $\int_X \mathbf{1}_A f \, d\mu = \int_X \mathbf{1}_A g \, d\mu$, for all $A \in \mathcal{A}$. Then $f = g$ μ-a.e.*

Proof. For all $n \in \mathbb{N}$, let $A_n \triangleq \{x \in X : f(x) - g(x) > \frac{1}{n}\}$. Then $\int_X \mathbf{1}_{A_n} (f - g) \, d\mu = 0$, so that $\mu(A_n) = 0$. Hence $\mu(\{x \in X : f(x) > g(x)\} = \mu(\bigcup_{n\in\mathbb{N}} A_n) = 0$. Similarly, $\mu(\{x \in X : f(x) < g(x)\} = 0$. Hence $f = g$ μ-a.e. □

The next result is the Radon-Nikodym theorem. For that, the definition of absolute continuity is needed.

Definition A.2.12. Let ν and μ be two measures on the same measurable space. Then ν is *absolutely continuous* with respect to μ if $\nu(A) = 0$, whenever $\mu(A) = 0$.

Proposition A.2.12. *(Radon-Nikodym theorem) Let $(X, \mathcal{A}, \mu)$ be a σ-finite measure space and ν a finite measure on $\mathcal{A}$ that is absolutely continuous with respect to μ. Then there exists a non-negative, integrable function $\breve{\nu} : X \to \mathbb{R}$ such that*

$$\nu(A) = \int_X \mathbf{1}_A \breve{\nu} \, d\mu,$$

for all $A \in \mathcal{A}$. Any two such $\breve{\nu}$ are equal μ-a.e.

Proof. See [43, p.175]. Uniqueness follows directly from Proposition A.2.11. □

The function $\breve{\nu}$ in Proposition A.2.12 is called the *Radon-Nikodym derivative* of ν with respect to μ.

A measurable function defined on a measure space induces a measure on its codomain.

Proposition A.2.13. *Let $(X, \mathcal{A}, \mu)$ be a measure space, $(Y, \mathcal{B})$ a measurable space, and $f : X \to Y$ a measurable function. Then $\mu \circ f^{-1}$ is a measure on $\mathcal{B}$, so that $(Y, \mathcal{B}, \mu \circ f^{-1})$ is a measure space. If $(X, \mathcal{A}, \mu)$ is a probability space, then so is $(Y, \mathcal{B}, \mu \circ f^{-1})$.*

Proof. The countable additivity of $\mu \circ f^{-1}$ follows from that for μ and the fact that f^{-1} preserves unions and intersections. The second part is obvious. □

Proposition A.2.14. *Let $(X, \mathcal{A}, \mu)$ be a measure space, $(Y, \mathcal{B})$ a measurable space, and $f : X \to Y$ and $g : Y \to \mathbb{R}$ measurable functions. Then*

$$\int_X (g \circ f) \, d\mu = \int_Y g \, d(\mu \circ f^{-1}),$$

whenever either side exists.

Proof. See [83, Lemma 1.22]. □

Let $(X, \mathcal{A}, \mu)$ be a measure space and Y a measurable subset of X. The *restriction* of μ to Y is the function $\mu_Y : \mathcal{A} \to \mathbb{R}$ defined by $\mu_Y(A) = \mu(A \cap Y)$, for all $A \in \mathcal{A}$. Clearly, μ_Y is a measure on $\mathcal{A}$.

If $f : X \to \mathbb{R}$ is measurable, then

$$\int_X f \, d\mu_Y = \int_X f \mathbf{1}_Y \, d\mu,$$

where $\mathbf{1}_Y$ is the indicator function for Y.

Measures on product spaces will be needed.

Definition A.2.13. Let I be an index set, and $(X_i, \mathcal{A}_i)$ a measurable space, for $i \in I$. A *cylinder* is a set of the form $\prod_{i \in I} B_i$, where $B_i \in \mathcal{A}_i$, for $i \in I$, and $B_i = X_i$, for all but finitely many $i \in I$.

Let $(X_i, \mathcal{A}_i, \mu_i)$ be a σ-finite measure space, for $i = 1, \ldots, n$. The *product measure* $\bigotimes_{i=1}^n \mu_i$ on $\prod_{i=1}^n X_i$ is defined first on cylinders by $(\bigotimes_{i=1}^n \mu_i)(A_1 \times \cdots \times A_n) = \mu_1(A_1) \ldots \mu_n(A_n)$, where $A_i \in \mathcal{A}_i$, for $1 = 1, \ldots, n$, and then extended to the σ-algebra $\bigotimes_{i=1}^n \mathcal{A}_i$ in a standard way [43, p.134]. Thus $(\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i, \bigotimes_{i=1}^n \mu_i)$ is a measure space.

Notation. Let $S \triangleq \{i_1, \ldots, i_m\} \subseteq \{1, \ldots, n\}$, where $i_1 < \cdots < i_m$. Then x_S denotes $(x_{i_1}, \ldots, x_{i_m})$. Also S^c denotes $\{1, \ldots, n\} \setminus S$.

Proposition A.2.15. *Let $(X_i, \mathcal{A}_i, \mu_i)$ be a σ-finite measure space, for $i = 1, \ldots, n$. If $f : \prod_{i=1}^n X_i \to \mathbb{R}$ is a non-negative measurable function and $M \subseteq \{1, \ldots, n\}$, then*

$$\lambda x_M . \int_{\prod_{p \in M^c} X_p} \lambda x_{M^c} . f(x_1, \ldots, x_n) \, d \bigotimes_{p \in M^c} \mu_p : \prod_{j \in M} X_j \to \mathbb{R}$$

is measurable.

Proof. See [83, Lemma 1.26]. (EXTEND TO INFINITE PRODUCT KERNELS) □

Here is the key result about integrating with respect to a product measure.

Proposition A.2.16. *(Fubini theorem) Let $(X_i, \mathcal{A}_i, \mu_i)$ be a σ-finite measure space, for $i = 1, \ldots, n$. If $f : \prod_{i=1}^n X_i \to \mathbb{R}$ is a non-negative measurable function and $M \subseteq \{1, \ldots, n\}$, then*

$$\begin{aligned} &\int_{\prod_{i=1}^n X_i} f \, d \bigotimes_{i=1}^n \mu_i \\ &= \int_{\prod_{j \in M} X_j} \left(\lambda x_M . \int_{\prod_{p \in M^c} X_p} \lambda x_{M^c} . f(x_1, \ldots, x_n) \, d \bigotimes_{p \in M^c} \mu_p \right) \, d \bigotimes_{j \in M} \mu_j . \end{aligned}$$

If $f : \prod_{i=1}^n X_i \to \mathbb{R}$ is a non-negative measurable function, then

$$\begin{aligned} &\int_{\prod_{i=1}^n X_i} f \, d \bigotimes_{i=1}^n \mu_i \\ &= \int_{X_n} \left(\lambda x_n . \int_{X_{n-1}} \left(\lambda x_{n-1} . \ \ldots \ \int_{X_1} \lambda x_1 . f(x_1, \ldots, x_n) \, d\mu_1 \ \ldots \right) \, d\mu_{n-1} \right) \, d\mu_n . \end{aligned}$$

More generally, if σ is a permuation of $\{1,\dots,n\}$, then

$$\begin{aligned}&\int_{\prod_{i=1}^n X_i} f \; d\bigotimes_{i=1}^n \mu_i \\ &= \int_{X_{\sigma(n)}} \left(\lambda x_{\sigma(n)}. \int_{X_{\sigma(n-1)}} \left(\lambda x_{\sigma(n-1)}. \; \dots \; \int_{X_{\sigma(1)}} \lambda x_{\sigma(1)}. f(x_1,\dots,x_n) \; d\mu_{\sigma(1)} \; \dots \right) \; d\mu_{\sigma(n-1)} \right) \; d\mu_{\sigma(n)}.\end{aligned}$$

Proof. See [83, Theorem 1.27], [43, Theorem 4.4.6], or [87, Theorem 14.16]. □

Probability kernels on product spaces can be formed naturally from probability kernels on each factor.

Proposition A.2.17. *Let $(X_i,\mathcal{A}_i)$ and $(Y_i,\mathcal{B}_i)$ be measurable spaces, and $\mu_i : X_i \to \mathcal{P}(Y_i)$ a probability kernel, for $i = 1,\dots,n$. Then*

$$\lambda(x_1,\dots,x_n).\bigotimes_{i=1}^n \mu_i(x_i) : \prod_{i=1}^n X_i \to \mathcal{P}(\prod_{i=1}^n Y_i)$$

is a probability kernel.

Proof. By Proposition A.2.4, it suffices to show that

$$\lambda(x_1,\dots,x_n).(\bigotimes_{i=1}^n \mu_i(x_i))(B) : \prod_{i=1}^n X_i \to \mathbb{R}$$

is measurable, for all $B \in \bigotimes_{i=1}^n \mathcal{B}_i$.

Let $\mathcal{P} \triangleq \{\prod_{i=1}^n B_i \mid B_i \in \mathcal{B}_i, \text{ for } i = 1,\dots,n\}$ and

$$\mathcal{L} \triangleq \{B \in \bigotimes_{i=1}^n \mathcal{B}_i \mid \lambda(x_1,\dots,x_n).(\bigotimes_{i=1}^n \mu_i(x_i))(B) : \prod_{i=1}^n X_i \to \mathbb{R} \text{ is measurable}\}.$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \bigotimes_{i=1}^n \mathcal{B}_i$.

Suppose that $\prod_{i=1}^n B_i \in \mathcal{P}$. Then

$$\begin{aligned}&\lambda(x_1,\dots,x_n).(\bigotimes_{i=1}^n \mu_i(x_i))(\prod_{i=1}^n B_i) \\ &= \lambda(x_1,\dots,x_n).\prod_{i=1}^n \mu_i(x_i)(B_i) \\ &= \lambda(y_1,\dots,y_n).\prod_{i=1}^n y_i \circ \lambda(x_1,\dots,x_n).(\mu_1(x_1)(B_1),\dots,\mu_n(x_n)(B_n))\end{aligned}$$

is measurable being a composition of the measurable functions

$$\lambda(y_1,\dots,y_n).\prod_{i=1}^n y_i : \mathbb{R}^n \to \mathbb{R} \text{ and}$$

$$\lambda(x_1,\dots,x_n).(\mu_1(x_1)(B_1),\dots,\mu_n(x_n)(B_n)) : \prod_{i=1}^n X_i \to \mathbb{R}^n.$$

Thus $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, $\prod_{i=1}^n X_i \in \mathcal{P}$ and $\mathcal{P} \subseteq \mathcal{L}$, so that $\prod_{i=1}^n X_i \in \mathcal{L}$.

Second, let $A, B \in \mathcal{L}$, where $A \subseteq B$. Then

$$\begin{aligned}&\lambda(x_1, \ldots, x_n).(\bigotimes_{i=1}^n \mu_i(x_i))(B \setminus A)\\ &= \lambda(x_1, \ldots, x_n).(\bigotimes_{i=1}^n \mu_i(x_i))(B) - \lambda(x_1, \ldots, x_n).(\bigotimes_{i=1}^n \mu_i(x_i))(A)\end{aligned}$$

is measurable being the difference of two measurable functions. Hence $B \setminus A \in \mathcal{L}$.

Third, let $(B_k)_{k\in\mathbb{N}}$ be an increasing sequence of sets in $\mathcal{L}$. Then

$$\begin{aligned}&\lambda(x_1, \ldots, x_n).(\bigotimes_{i=1}^n \mu_i(x_i))(\bigcup_{k\in\mathbb{N}} B_k)\\ &= \lim_{k\to\infty} \lambda(x_1, \ldots, x_n).(\bigotimes_{i=1}^n \mu_i(x_i))(B_k)\end{aligned}$$

is measurable being the limit of a sequence of real-valued measurable functions. Hence $\bigcup_{k\in\mathbb{N}} B_k \in \mathcal{L}$.

It now follows from the monotone-class theorem (Proposition A.1.2) that $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. Hence the result. □

Here is a useful relationship between product measures and fusions.

Proposition A.2.18. *Let $(X_i, \mathcal{A}_i)$ and $(Y_i, \mathcal{B}_i)$ be measurable spaces, $\mu_i : \mathcal{P}(X_i)$ a probability measure, and $\nu_i : X_i \to \mathcal{P}(Y_i)$ a probability kernel, for $i = 1, \ldots, n$. Then*

$$\bigotimes_{i=1}^n (\mu_i \odot \nu_i) = \bigotimes_{i=1}^n \mu_i \odot \lambda(x_1, \ldots, x_n).\bigotimes_{i=1}^n \nu_i(x_i).$$

Proof. Proposition A.2.17 shows that $\lambda(x_1, \ldots, x_n).\bigotimes_{i=1}^n \nu_i(x_i) : \prod_{i=1}^n X_i \to \mathcal{P}(\prod_{i=1}^n Y_i)$ is a probability kernel.

Let $\mathcal{P} \triangleq \{\prod_{i=1}^n B_i \mid B_i \in \mathcal{B}_i, \text{ for } i = 1, \ldots, n\}$ and

$$\mathcal{L} \triangleq \{B \in \bigotimes_{i=1}^n \mathcal{B}_i \mid (\bigotimes_{i=1}^n (\mu_i \odot \nu_i))(B) = (\bigotimes_{i=1}^n \mu_i \odot \lambda(x_1, \ldots, x_n).\bigotimes_{i=1}^n \nu_i(x_i))(B)\}.$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \bigotimes_{i=1}^n \mathcal{B}_i$.

Suppose that $\prod_{i=1}^n B_i \in \mathcal{P}$. Then

$$
\begin{aligned}
&(\bigotimes_{i=1}^n(\mu_i \odot \nu_i))(\prod_{i=1}^n B_i) \\
&= \prod_{i=1}^n(\mu_i \odot \nu_i)(B_i) \\
&= \prod_{i=1}^n \int_{X_i} \lambda x_i.\nu_i(x_i)(B_i)\, d\mu_i \\
&= \int_{\prod_{i=1}^n X_i} \lambda(x_1,\ldots,x_n).\prod_{i=1}^n \nu_i(x_i)(B_i)\, d\bigotimes_{i=1}^n \mu_i \qquad \text{[Proposition A.2.16]} \\
&= \int_{\prod_{i=1}^n X_i} \lambda(x_1,\ldots,x_n).(\bigotimes_{i=1}^n \nu_i(x_i))(\prod_{i=1}^n B_i)\, d\bigotimes_{i=1}^n \mu_i \\
&= (\bigotimes_{i=1}^n \mu_i \odot \lambda(x_1,\ldots,x_n).\bigotimes_{i=1}^n \nu_i(x_i))(\prod_{i=1}^n B_i).
\end{aligned}
$$

Thus $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, $\prod_{i=1}^n X_i \in \mathcal{P}$ and $\mathcal{P} \subseteq \mathcal{L}$, so that $\prod_{i=1}^n X_i \in \mathcal{L}$.

Second, let $A, B \in \mathcal{L}$, where $A \subseteq B$. Then

$$
\begin{aligned}
&(\bigotimes_{i=1}^n(\mu_i \odot \nu_i))(B \setminus A) \\
&= (\bigotimes_{i=1}^n(\mu_i \odot \nu_i))(B) - (\bigotimes_{i=1}^n(\mu_i \odot \nu_i))(A) \\
&= (\bigotimes_{i=1}^n \mu_i \odot \lambda(x_1,\ldots,x_n).\bigotimes_{i=1}^n \nu_i(x_i))(B) - (\bigotimes_{i=1}^n \mu_i \odot \lambda(x_1,\ldots,x_n).\bigotimes_{i=1}^n \nu_i(x_i))(A) \\
&= (\bigotimes_{i=1}^n \mu_i \odot \lambda(x_1,\ldots,x_n).\bigotimes_{i=1}^n \nu_i(x_i))(B \setminus A)
\end{aligned}
$$

Hence $B \setminus A \in \mathcal{L}$.

Third, let $(B_k)_{k\in\mathbb{N}}$ be an increasing sequence of sets in $\mathcal{L}$. Then

$$
\begin{aligned}
&(\bigotimes_{i=1}^n(\mu_i \odot \nu_i))(\bigcup_{k\in\mathbb{N}} B_k) \\
&= \lim_{k\to\infty} (\bigotimes_{i=1}^n(\mu_i \odot \nu_i))(B_k) \\
&= \lim_{k\to\infty} (\bigotimes_{i=1}^n \mu_i \odot \lambda(x_1,\ldots,x_n).\bigotimes_{i=1}^n \nu_i(x_i))(B_k) \\
&= (\bigotimes_{i=1}^n \mu_i \odot \lambda(x_1,\ldots,x_n).\bigotimes_{i=1}^n \nu_i(x_i))(\bigcup_{k\in\mathbb{N}} B_k).
\end{aligned}
$$

Hence $\bigcup_{k\in\mathbb{N}} B_k \in \mathcal{L}$.

It now follows from the monotone-class theorem (Proposition A.1.2) that $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. Hence the result. □

Let $(X_i, \mathcal{A}_i, \mu_i)$ be a measure space, for all $i \in I$. Define the *sum measure*

$$\bigoplus_{i\in I} \mu_i : \bigoplus_{i\in I} \mathcal{A}_i \to [0,\infty]$$

by

$$(\bigoplus_{i\in I} \mu_i)(\coprod_{i\in I} A_i) = \sum_{i\in I} \mu_i(A_i),$$

where $A_i \in \mathcal{A}_i$, for $i \in I$. (The sum on the right hand side is finite only if at most countably many $\mu_i(A_i)$ are non-zero.) Suppose that $(\coprod_{i\in I} A_i^{(j)})_{j=1}^{\infty}$ is a sequence of pairwise disjoint sets in $\bigoplus_{i\in I} \mathcal{A}_i$. Then

$$\begin{aligned}
&(\bigoplus_{i\in I} \mu_i)(\bigcup_{j=1}^{\infty} \coprod_{i\in I} A_i^{(j)}) \\
&= (\bigoplus_{i\in I} \mu_i)(\coprod_{i\in I} \bigcup_{j=1}^{\infty} A_i^{(j)}) \\
&= \sum_{i\in I} \mu_i(\bigcup_{j=1}^{\infty} A_i^{(j)}) \\
&= \sum_{i\in I} \sum_{j=1}^{\infty} \mu_i(A_i^{(j)}) \\
&= \sum_{j=1}^{\infty} \sum_{i\in I} \mu_i(A_i^{(j)}) \\
&= \sum_{j=1}^{\infty} (\bigoplus_{i\in I} \mu_i)(\coprod_{i\in I} A_i^{(j)}).
\end{aligned}$$

Thus $\bigoplus_{i\in I} \mu_i$ is a measure and so $(\coprod_{i\in I} X_i, \bigoplus_{i\in I} \mathcal{A}_i, \bigoplus_{i\in I} \mu_i)$ is a measure space. For example, if each μ_i is counting measure on X_i, then $\bigoplus_{i\in I} \mu_i$ is counting measure on $\coprod_{i\in I} X_i$. If $f : \coprod_{i\in I} X_i \to \mathbb{R}$ is a non-negative measurable function, then

$$\int_{\coprod_{i\in I} X_i} f \, d \bigoplus_{i\in I} \mu_i = \sum_{i\in I} \int_{X_i} f|_{X_i} \, d\mu_i.$$

Let $(X, \mathcal{A}, \mu)$ be a probability space and Y a measurable subset of X such that $\mu(Y) > 0$. The *normalized restriction* of μ to Y is the function $\mu|_Y : \mathcal{A}|_Y \to \mathbb{R}$ defined by

$$\mu|_Y(A \cap Y) = \frac{\mu(A \cap Y)}{\mu(Y)},$$

for all $A \in \mathcal{A}$. Clearly, $\mu|_Y$ is a probability measure on $\mathcal{A}|_Y$, so that $(Y, \mathcal{A}|_Y, \mu|_Y)$ is a probability space. If $f : Y \to \mathbb{R}$ is measurable, then

$$\int_Y f \, d\mu|_Y = \frac{\int_X \bar{f} \, d\mu}{\mu(Y)},$$

where $\bar{f}$ is f extended to X by defining it to be 0 on $X \setminus Y$.

The concept of a projective product will be needed in Bayes theorem for probability kernels below.

Definition A.2.14. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $f : X \to Y \to \mathbb{R}$ a non-negative function such that $\lambda(x, y).f(x)(y) : X \times Y \to \mathbb{R}$ is measurable, and $\mu : X \to \mathcal{P}(Y)$ a probability kernel such that $0 < \int_Y f(x) \, d\mu(x) < \infty$, for all $x \in X$. Then the *projective product* $f * \mu : X \to \mathcal{P}(Y)$ *of* f *and* μ is defined by

$$(f * \mu)(x)(B) = \frac{\int_Y \mathbf{1}_B f(x) \, d\mu(x)}{\int_Y f(x) \, d\mu(x)},$$

for all $x \in X$ and $B \in \mathcal{B}$.

Taking X to be a singleton set in Definition A.2.14, the following special case is obtained. Let $(Y, \mathcal{B})$ be a measurable space, $f : Y \to \mathbb{R}$ a non-negative measurable function, and $\mu : \mathcal{P}(Y)$ a probability measure such that $0 < \int_Y f \, d\mu < \infty$. Then the projective product $f * \mu : \mathcal{P}(Y)$ of f and μ is defined by

$$(f * \mu)(B) = \frac{\int_Y \mathbf{1}_B f \, d\mu}{\int_Y f \, d\mu},$$

for all $B \in \mathcal{B}$.

Note that if $f : X \to Y \to \mathbb{R}$ has the property that $f(x)$ is a (possibly different) constant function, for all $x \in X$, then $f * \mu = \mu$. In the case when X is a singleton set, if $f : Y \to \mathbb{R}$ is a constant function, then $f * \mu = \mu$.

Proposition A.2.19. *Let* $(X, \mathcal{A})$ *and* $(Y, \mathcal{B})$ *be measurable spaces,* $f : X \to Y \to \mathbb{R}$ *a non-negative function such that* $\lambda(x, y).f(x)(y) : X \times Y \to \mathbb{R}$ *is measurable, and* $\mu : X \to \mathcal{P}(Y)$ *a probability kernel such that* $0 < \int_Y f(x) \, d\mu(x) < \infty$*, for all* $x \in X$*. Then* $f * \mu : X \to \mathcal{P}(Y)$ *is a probability kernel.*

Proof. For all $x \in X$, $f(x)$ is measurable, $(f * \mu)(x)(Y) = 1$, and, by the monotone converergence theorem (Proposition A.2.2), $(f * \mu)(x)$ is countably additive. Thus $f * \mu : X \to \mathcal{P}(Y)$ is well-defined. By Proposition A.2.6, $f * \mu$ is measurable. □

Note that, for all $x \in X$, $(f * \mu)(x)$ is a probability measure that is absolutely continuous with respect to $\mu(x)$.

Example A.2.10. Let $(Y, \mathcal{B})$ be a measurable space, $y_0 \in Y$ such that $\{y_0\} \in \mathcal{B}$, $\mu : \mathcal{P}(Y)$ a probability measure such that $0 < \mu(\{y_0\}) < \infty$, and $K > 0$. Define $f : Y \to \mathbb{R}$ by

$$f(y) = \begin{cases} K & \text{if } y = y_0 \\ 0 & \text{otherwise,} \end{cases}$$

for all $y \in Y$. Clearly, f is measurable and $0 < \int_Y f \, d\mu < \infty$. For all $B \in \mathcal{B}$,

$$\begin{aligned} & (f * \mu)(B) \\ = & \frac{\int_Y \mathbf{1}_B f \, d\mu}{\int_Y f \, d\mu} \\ = & \begin{cases} 1 & \text{if } y_0 \in B \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Hence $f * \mu = \delta_{y_0}$.

Proposition A.2.20. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $f : X \to Y \to \mathbb{R}$ a non-negative function such that $\lambda(x, y).f(x)(y) : X \times Y \to \mathbb{R}$ is measurable, and $\mu : X \to \mathcal{P}(Y)$ a probability kernel such that $0 < \int_Y f(x) \, d\mu(x) < \infty$, for all $x \in X$. Let $g : Y \to \mathbb{R}$ be a function that is integrable with respect to $(f * \mu)(x)$, for all $x \in X$. Then, for all $x \in X$,*

$$\int_Y g \, d(f * \mu)(x) = \frac{\int_Y gf(x) \, d\mu(x)}{\int_Y f(x) \, d\mu(x)}.$$

Proof. Let g be an indicator function $\mathbf{1}_B$, where $B \in \mathcal{B}$. Then, for all $x \in X$,

$$\int_Y \mathbf{1}_B \, d(f * \mu)(x) = (f * \mu)(x)(B) = \frac{\int_Y \mathbf{1}_B f(x) \, d\mu(x)}{\int_Y f(x) \, d\mu(x)}.$$

Hence the result holds for measurable indicator functions. By linearity of the integral, it holds for simple functions. By Proposition A.2.1 and the monotone convergence theorem (Proposition A.2.2), it also holds for non-negative measurable functions g. Then, since g is integrable with respect to $(f * \mu)(x)$, for all $x \in X$, the result follows. □

It will be convenient for later applications to extend the definition of projective product.

Definition A.2.15. Let I be a finite index set, $J \subseteq I$, $(X_i, \mathcal{A}_i)$, for all $i \in I$, and $(Y, \mathcal{B})$ measurable spaces, $f : \prod_{j \in J} X_j \to Y \to \mathbb{R}$ a non-negative function such that the mapping from $\prod_{j \in J} X_j \times Y$ to $\mathbb{R}$ defined by $((x_j)_{j \in J}, y) \mapsto f((x_j)_{j \in J})(y)$ is measurable, and $\mu : \prod_{i \in I} X_i \to \mathcal{P}(Y)$ a probability kernel such that $0 < \int_Y f((x_j)_{j \in J}) \, d\mu((x_i)_{i \in I}) < \infty$, for all $(x_i)_{i \in I} \in \prod_{i \in I} X_i$. Then the *projective product* $f * \mu : \prod_{i \in I} X_i \to \mathcal{P}(Y)$ *of* f *and* μ is defined to be

$$f * \mu = \lambda((x_i)_{i \in I}).f((x_j)_{j \in J}) * \mu.$$

Here is a useful relationship between product measures and projective products.

Proposition A.2.21. *Let $(X_i, \mathcal{A}_i)$ be a measurable space, $\mu_i : \mathcal{P}(X_i)$ a probability measure, and $f_i : X_i \to \mathbb{R}$ a non-negative measurable function such that $0 < \int_{X_i} f_i \, d\mu < \infty$, for $i = 1, \ldots, n$. Then*

$$\bigotimes_{i=1}^{n} (f_i * \mu_i) = \lambda(x_1, \ldots, x_n). \prod_{i=1}^{n} f_i(x_i) \; * \bigotimes_{i=1}^{n} \mu_i.$$

Proof. Note that $\lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i) * \bigotimes_{i=1}^n \mu_i$ is well-defined since

$$0 < \int_{\prod_{i=1}^n X_i} \lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i)\; d\bigotimes_{i=1}^n \mu_i = \prod_{i=1}^n \int_{X_i} f_i\; d\mu_i < \infty,$$

by Fubini's theorem (Proposition A.2.16).

Let $\mathcal{P} \triangleq \{\prod_{i=1}^n A_i \mid A_i \in \mathcal{A}_i, \text{ for } i = 1,\dots,n\}$ and

$$\mathcal{L} \triangleq \{A \in \bigotimes_{i=1}^n \mathcal{A}_i \mid \bigotimes_{i=1}^n (f_i * \mu_i)(A) = (\lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i) * \bigotimes_{i=1}^n \mu_i)(A)\}.$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \bigotimes_{i=1}^n \mathcal{A}_i$.

Suppose that $\prod_{i=1}^n A_i \in \mathcal{P}$. Then

$$\begin{aligned}
&\bigotimes_{i=1}^n (f_i * \mu_i)(\prod_{i=1}^n A_i) \\
&= \prod_{i=1}^n (f_i * \mu_i)(A_i) \\
&= \prod_{i=1}^n \frac{\int_{X_i} \mathbf{1}_{A_i} f_i\; d\mu_i}{\int_{X_i} f_i\; d\mu_i} \\
&= \frac{\int_{\prod_{i=1}^n X_i} \mathbf{1}_{\prod_{i=1}^n A_i}\, \lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i)\; d\bigotimes_{i=1}^n \mu_i}{\int_{\prod_{i=1}^n X_i} \lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i)\; d\bigotimes_{i=1}^n \mu_i} \qquad \text{[Proposition A.2.16]} \\
&= (\lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i) * \bigotimes_{i=1}^n \mu_i)(\prod_{i=1}^n A_i).
\end{aligned}$$

Thus $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, $\prod_{i=1}^n X_i \in \mathcal{P}$ and $\mathcal{P} \subseteq \mathcal{L}$, so that $\prod_{i=1}^n X_i \in \mathcal{L}$.

Second, let $A, B \in \mathcal{L}$, where $A \subseteq B$. Then

$$\begin{aligned}
&\bigotimes_{i=1}^n (f_i * \mu_i)(B \setminus A) \\
&= \bigotimes_{i=1}^n (f_i * \mu_i)(B) - \bigotimes_{i=1}^n (f_i * \mu_i)(A) \\
&= (\lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i) * \bigotimes_{i=1}^n \mu_i)(B) - (\lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i) * \bigotimes_{i=1}^n \mu_i)(A) \\
&= (\lambda(x_1,\dots,x_n).\prod_{i=1}^n f_i(x_i) * \bigotimes_{i=1}^n \mu_i)(B \setminus A).
\end{aligned}$$

Hence $B \setminus A \in \mathcal{L}$.

Third, let $(A_k)_{k\in\mathbb{N}}$ be an increasing sequence of sets in $\mathcal{L}$. Then

$$\begin{aligned}
&\bigotimes_{i=1}^{n}(f_i * \mu_i)(\bigcup_{k\in\mathbb{N}} A_k)\\
&= \lim_{k\to\infty} \bigotimes_{i=1}^{n}(f_i * \mu_i)(A_k)\\
&= \lim_{k\to\infty} (\lambda(x_1,\ldots,x_n).\prod_{i=1}^{n} f_i(x_i) * \bigotimes_{i=1}^{n}\mu_i)(A_k)\\
&= (\lambda(x_1,\ldots,x_n).\prod_{i=1}^{n} f_i(x_i) * \bigotimes_{i=1}^{n}\mu_i)(\bigcup_{k\in\mathbb{N}} A_k).
\end{aligned}$$

Hence $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{L}$.

It now follows from the monotone-class theorem (Proposition A.1.2) that $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. Hence the result. □

A.3 Densities and Conditional Densities

In applications, the most common way of defining probability measures (resp., probability kernels) is by means of densities (conditional densities).

Definition A.3.1. Let $(X, \mathcal{A}, \nu)$ be a measure space and $h : X \to \mathbb{R}$ a measurable function. Then h is a *density (on $(X, \mathcal{A}, \nu)$)* if (i) $h(x) \geq 0$, for all $x \in X$, and (ii) $\int_X h\, d\nu = 1$.

A density is often called a *probability density function* (and abbreviated to *pdf*) in the literature. In the following, the phrase 'h is a density on $(X, \mathcal{A}, \nu)$' is abbreviated to 'h is a density on X', with $\mathcal{A}$ and ν understood.

Notation. With the same understanding, $\mathcal{D}(X)$ denotes the set of densities on $(X, \mathcal{A}, \nu)$.

$\mathcal{D}(X)$ is a measurable space with the σ-algebra inherited from the σ-algebra on $\mathbb{R}^X$ given by the evaluation maps. (See Definition A.1.14.)

Let $(X, \mathcal{A}, \mu)$ be a measure space, f a density on $(X, \mathcal{A}, \mu)$, and $Y \in \mathcal{A}$. Assuming that $\int_Y f\, d\mu \neq 0$, define $f|_Y : Y \to \mathbb{R}$ by

$$f|_Y(y) = \frac{f(y)}{\int_Y f\, d\mu},$$

for all $y \in Y$. Clearly, $f|_Y$ is a density on $(Y, \mathcal{A}|_Y, \mu|_Y)$.

Densities induce probability measures, in a natural way.

Definition A.3.2. Let $(X, \mathcal{A}, \nu)$ be a measure space and $h : X \to \mathbb{R}$ a non-negative measurable function. Define $h \cdot \nu : \mathcal{A} \to \mathbb{R}$ by

$$(h \cdot \nu)(A) = \int_X \mathbf{1}_A h\, d\nu,$$

for all $A \in \mathcal{A}$.

Proposition A.3.1. *Let $(X, \mathcal{A}, \nu)$ be a measure space and $h : X \to \mathbb{R}$ a non-negative measurable function. Then $h \cdot \nu$ is a measure. Furthermore, if h is a density, then $h \cdot \nu$ is a probability measure.*

Proof. See [83, p.12]. □

Example A.3.1. Define $h : \mathbb{N}_0 \to \mathbb{R}$ by $h(n) = e^{-\lambda}\frac{\lambda^n}{n!}$, for some $\lambda > 0$. Then h is the *Poisson density* with parameter λ. The associated probability measure on the measurable space $(\mathbb{N}_0, \mathbb{B}^{\mathbb{N}_0})$ is $h \cdot c$, where c is counting measure. Thus $(h \cdot c)(A) = \int_{\mathbb{N}_0} \mathbf{1}_A h \; dc = \sum_{n \in A} h(n)$, for all $A \subseteq \mathbb{N}_0$.

Example A.3.2. Let $(X, 2^X, c)$ be a measure space, where $X \triangleq \{x_1, \ldots, x_n\}$ is a finite set and c is counting measure. Let $(a_i)_{i=1}^n$ be a sequence of non-negative real numbers such that $\sum_{i=1}^n a_i = 1$. The *categorical density* $f : \mathcal{D}(X)$ (determined by $(a_i)_{i=1}^n$) is defined by $f(x_i) = a_i$, for $i = 1, \ldots, n$. Note that $\int_X f \; dc = \sum_{i=1}^n f(x_i) = 1$. The categorical measure μ determined by $(a_i)_{i=1}^n$ is given by $f \cdot c$, since $(f \cdot c)(A) = \int_X \mathbf{1}_A f \; dc = \sum_{x_i \in A} a_i = \mu(A)$, for all $A \subseteq X$. Categorical densities are usually called probability mass functions [163].

Example A.3.3. The *uniform density* $f : \mathcal{D}(\mathbb{R})$ on $[a, b]$ is defined by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

The uniform density on $[a, b]$ is denoted by $\mathcal{U}(a, b)$.

Example A.3.4. The *Gaussian density* $f : \mathcal{D}(\mathbb{R})$ with mean μ and variance σ^2 is defined by

$$f = \lambda x. \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

More generally, the *Gaussian density* $f : \mathcal{D}(\mathbb{R}^m)$, where $m \in \mathbb{N}$, with mean μ and covariance Σ is defined by

$$f = \lambda x. \frac{1}{(2\pi)^{m/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}.$$

Here, $\mu \in \mathbb{R}^m$, Σ is a symmetric, positive definite, real, $m \times m$ matrix, $|\Sigma|$ is the determinant of Σ, and a superscript T denotes transpose. (In the matrix expression in the exponent, μ and x are regarded as column vectors.) Let

$$\mathcal{N}(\mu, \Sigma) \triangleq \lambda x. \frac{1}{(2\pi)^{m/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}.$$

The importance of the Radon-Nikodym theorem (Proposition A.2.12) is that it shows that, under weak conditions, primarily that of absolute continuity, one can work with densities rather than probability measures, which is generally done in applications. However, this is not always possible. Consider the σ-finite measure space $(\mathbb{R}, \mathcal{B}, \lambda)$, where $\mathcal{B}$ is the Borel σ-algebra and λ is Lebesgue measure. The Dirac measure δ_a, for any fixed $a \in \mathbb{R}$, is not absolutely continuous with respect to λ, and there does not exist a density $h \in \mathcal{D}(\mathbb{R})$ such that $\delta_a = h \cdot \lambda$. But Dirac measures are useful in practice.

Here is a situation in which a Dirac measure does have a density.

Proposition A.3.2. *Let $(X, \mathcal{A}, \mu)$ be a measure space and x_0 a fixed element of X such that $\{x_0\} \in \mathcal{A}$ and $0 < \mu(\{x_0\}) < \infty$. Define $h : X \to \mathbb{R}$ by*

$$h(x) = \begin{cases} 1/\mu(\{x_0\}) & \text{if } x = x_0 \\ 0 & \text{otherwise,} \end{cases}$$

for all $x \in X$. Then $h \cdot \mu = \delta_{x_0}$.

Proof. Clearly $h \in \mathcal{D}(X)$. For all $A \in \mathcal{A}$,

$$\begin{aligned} & (h \cdot \mu)(A) \\ = & \int_X \mathbf{1}_A h \, d\mu \\ = & \begin{cases} 1 & \text{if } x_0 \in A \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Hence $h \cdot \mu = \delta_{x_0}$. □

Since δ_{x_0} is absolutely continuous with respect to μ, the Radon-Nykodym theorem shows that there exists an h such that $h \cdot \mu = \delta_{x_0}$. The point of Proposition A.3.2 is to construct such an h. If X is countable, $\mathcal{A}$ is 2^X, and μ is counting measure, then Proposition A.3.2 is applicable to any element x_0 of X. See also Proposition A.3.5 below.

Proposition A.3.3. *Let $(X, \mathcal{A}, \nu)$ be a measure space, $h : X \to \mathbb{R}$ a non-negative measurable function, and $g : X \to \mathbb{R}$ a measurable function. Then g is $h \cdot \nu$-integrable iff gh is ν-integrable. In this case,*

$$\int_X g \, d(h \cdot \nu) = \int_X gh \, d\nu.$$

Proof. See [83, Lemma 1.23] or [87, Theorem 4.15]. □

The product of densities can also be defined.

Definition A.3.3. Let $(X_i, \mathcal{A}_i, \nu_i)$ be a measure space and $h_i : X_i \to \mathbb{R}$ a density, for $i = 1, \ldots, n$. Then the *product density* $\bigotimes_{i=1}^n h_i : \prod_{i=1}^n X_i \to \mathbb{R}$ (on $(\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i, \bigotimes_{i=1}^n \nu_i)$) is defined by

$$\bigotimes_{i=1}^n h_i = \lambda(x_1, \ldots, x_n).\prod_{i=1}^n h_i(x_i).$$

Clearly, $\bigotimes_{i=1}^n h_i$ is indeed a density, by Proposition A.2.16.

Conditional densities will also be needed.

Definition A.3.4. Let $(X, \mathcal{A})$ be a measurable space and $(Y, \mathcal{B}, \nu)$ a measure space. A *conditional density* is a function $h : X \to \mathcal{D}(Y)$ such that $\lambda(x, y).h(x)(y) : X \times Y \to \mathbb{R}$ is measurable.

Note that a conditional density is, in effect, *jointly* measurable (in X and Y). This is strictly stronger than just requiring that a conditional density be measurable as a mapping from X to $\mathcal{D}(Y)$, which is equivalent to *separate* measurability (in X and in Y).

Example A.3.5. A *linear Gaussian conditional density*

$$h : \mathbb{R}^n \to \mathcal{D}(\mathbb{R}),$$

where $n \in \mathbb{N}$, is defined by

$$h(x_1, \ldots, x_n) = \lambda x. \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(x - \left(\sum_{i=1}^{n} w_i x_i + b \right) \right)^2 \right\},$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$. Here, $w_1, \ldots, w_n$ and b are parameters giving the mean and σ^2 is the variance.

A *noisy-OR conditional density*

$$h : \mathbb{B}^n \to \mathcal{D}(\mathbb{B}),$$

where $n \in \mathbb{N}$, is defined by

$$\begin{aligned} h(b_1, \ldots, b_n)(\mathsf{F}) &= \lambda_0 \prod_{i \,:\, b_i = \mathsf{T}} \lambda_i \\ h(b_1, \ldots, b_n)(\mathsf{T}) &= 1 - \lambda_0 \prod_{i \,:\, b_i = \mathsf{T}} \lambda_i, \end{aligned}$$

for all $(b_1, \ldots, b_n) \in \mathbb{B}^n$. Here, $\lambda_i \in (0, 1]$ is a parameter, for $i = 0, \ldots, n$.

Example A.3.6. This example considers densities on $\mathbb{B}^m$, for $m \in \mathbb{N}$.

Let $f_i : \mathcal{D}(\mathbb{B})$, for $i = 1, \ldots, m$. Each f_i has a single parameter that determines $f_i(\mathsf{T})$ (or $f_i(\mathsf{F})$). Define $f : \mathcal{D}(\mathbb{B}^m)$ by

$$f(b_1, \ldots, b_m) = \prod_{i=1}^{m} f_i(b_i),$$

for all $(b_1, \ldots, b_m) \in \mathbb{B}^m$. The definition of f requires m parameters in all.

Next suppose that the conditional density $g_i : \mathbb{B}^{i-1} \to \mathcal{D}(\mathbb{B})$, for $i = 1, \ldots, m$, requires the maximum number of parameters, 2^{i-1}, for its definition. Define $g : \mathcal{D}(\mathbb{B}^m)$ by

$$g(b_1, \ldots, b_m) = \prod_{i=1}^{m} g_i(b_1, \ldots, b_{i-1})(b_i),$$

for all $(b_1, \ldots, b_m) \in \mathbb{B}^m$. The definition of g requires $2^m - 1$ parameters in all.

Conditional densities induce probability kernels, in a natural way.

Definition A.3.5. Let $(X, \mathcal{A})$ be a measurable space, $(Y, \mathcal{B}, \nu)$ a measure space, and $h : X \to \mathcal{D}(Y)$ a conditional density. Define $h \cdot \nu : X \to \mathcal{P}(Y)$ by

$$(h \cdot \nu)(x) = \lambda B. \int_Y \mathbf{1}_B \, h(x) \, d\nu,$$

for all $x \in X$.

Put another way, $h \cdot \nu : X \to \mathcal{P}(Y)$ is defined by $(h \cdot \nu)(x) = h(x) \cdot \nu$, for all $x \in X$.

Proposition A.3.4. *Let $(X, \mathcal{A})$ be a measurable space, $(Y, \mathcal{B}, \nu)$ a measure space, and $h : X \to \mathcal{D}(Y)$ a conditional density. Then $h \cdot \nu$ is a probability kernel.*

Proof. See [24, p.37, p.46] or Proposition A.2.6. □

Generally, Dirac measures do not have densities, since they usually are not absolutely continuous with respect to the underlying measure. This can be inconvenient in applications. However, in some situations, the Dirac measures only appear in a fusion with a probability kernel that has a conditional density. The next proposition shows that the fusion of a Dirac measure and a probability kernel with a conditional density always has a density.

Proposition A.3.5. *Let $(X_1, \mathcal{A}_1)$ be a measurable space, $(X_2, \mathcal{A}_2, \nu_2)$ a measure space, $\mu_1 : \mathcal{P}(X_1)$ a probability measure, and $\mu_2 : X_1 \to \mathcal{P}(X_2)$ a probability kernel. Suppose that $\mu_1 = \sum_{j=1}^{m} c_j \delta_{a_j}$, where $c_j \geq 0$ and $a_j \in X_1$, for $j = 1, \ldots, m$, and $\sum_{j=1}^{m} c_j = 1$. Let $h_2 : X_1 \to \mathcal{D}(X_2)$ be a conditional density such that $\mu_2 = h_2 \cdot \nu_2$. Then*

$$\mu_1 \odot \mu_2 = \sum_{j=1}^{m} c_j \, \mu_2(a_j) = (\sum_{j=1}^{m} c_j h_2(a_j)) \cdot \nu_2.$$

Proof.

$$\begin{aligned}
&\mu_1 \odot \mu_2 \\
&= (\sum_{j=1}^{m} c_j \delta_{a_j}) \odot \mu_2 \\
&= \lambda A_2. \int_{X_2} \lambda x_1. \mu_2(x_1)(A_2) \; d \sum_{j=1}^{m} c_j \delta_{a_j} \\
&= \lambda A_2. \sum_{j=1}^{m} c_j \, \mu_2(a_j)(A_2) \\
&= \sum_{j=1}^{m} c_j \, \mu_2(a_j) \\
&= \sum_{j=1}^{m} c_j \, (h_2(a_j) \cdot \nu_2) \\
&= (\sum_{j=1}^{m} c_j \, h_2(a_j)) \cdot \nu_2.
\end{aligned}$$

□

A generalization of the concept of 'almost everywhere' for the equality of conditional densities will be useful.

Definition A.3.6. Let $(X, \mathcal{A})$ be a measurable space, $(Y, \mathcal{B}, \nu)$ a measure space, and $f : X \to \mathcal{D}(Y)$ and $g : X \to \mathcal{D}(Y)$ conditional densities. Then $f = g$ ν-a.e. if, for all $x \in X$, $f(x) = g(x)$ ν-a.e.

Example A.3.7. Here are more examples of probability kernels useful for applications.

A *linear Gaussian probability kernel* $\mu : \mathbb{R}^n \to \mathcal{P}(\mathbb{R})$ has the form $\mu = h \cdot \lambda$, where λ is Lebesgue measure on $\mathbb{R}$ and

$$h : \mathbb{R}^n \to \mathcal{D}(\mathbb{R})$$

is a linear Gaussian conditional density.

A *noisy-OR probability kernel* $\mu : \mathbb{B}^n \to \mathcal{P}(\mathbb{B})$ has the form $\mu = h \cdot c$, where c is counting measure on $\mathbb{B}$ and

$$h : \mathbb{B}^n \to \mathcal{D}(\mathbb{B})$$

is a noisy-OR conditional density.

If $\mu : X \to \mathcal{P}(Y)$ is a probability kernel and there exists $h : X \to \mathcal{D}(Y)$ such that $\mu = h \cdot \nu$, then $\mu(x)$ is absolutely continuous with respect to ν, for all $x \in X$. Thus absolute continuity is a *necessary* condition for the existence of h.

Proposition A.3.6. *Let $(X, \mathcal{A})$ be a measurable space, $(Y, \mathcal{B}, \nu)$ a measure space, $h : X \to \mathcal{D}(Y)$ a conditional density, and $g : X \to Y \to \mathbb{R}$, where $g(x)$ is a measurable function, for all $x \in X$. Then, for all $x \in X$, $g(x)$ is $(h \cdot \nu)(x)$-integrable iff $g(x)h(x)$ is ν-integrable. In this case, for all $x \in X$,*

$$\int_Y g(x)\ d(h \cdot \nu)(x) = \int_Y g(x)h(x)\ d\nu.$$

Proof. The result is a direct consequence of Proposition A.3.3. □

The following definition gives the fusion of two conditional densities.

Definition A.3.7. Let $(X_0, \mathcal{A}_0)$ be a measurable space, $(X_1, \mathcal{A}_1, \nu_1)$ and $(X_2, \mathcal{A}_2, \nu_2)$ σ-finite measure spaces, and $h_1 : X_0 \to \mathcal{D}(X_1)$ and $h_2 : X_0 \times X_1 \to \mathcal{D}(X_2)$ conditional densities. The *fusion* $h_1 \odot h_2 : X_0 \to \mathcal{D}(X_2)$ of h_1 and h_2 is defined by

$$(h_1 \odot h_2)(x_0) = \lambda x_2. \int_{X_1} \lambda x_1. h_2(x_0, x_1)(x_2)\ h_1(x_0)\ d\nu_1,$$

for all $x_0 \in X_0$.

Proposition A.3.7. *Let $(X_0, \mathcal{A}_0)$ be a measurable space, $(X_1, \mathcal{A}_1, \nu_1)$ and $(X_2, \mathcal{A}_2, \nu_2)$ σ-finite measure spaces, and $h_1 : X_0 \to \mathcal{D}(X_1)$ and $h_2 : X_0 \times X_1 \to \mathcal{D}(X_2)$ conditional densities. Then $h_1 \odot h_2 : X_0 \to \mathcal{D}(X_2)$ is a conditional density.*

Proof. Note first that $\lambda(x_0, x_1, x_2).h_2(x_0, x_1)(x_2)\, h_1(x_0)(x_1) : X_0 \times X_1 \times X_2 \to \mathbb{R}$ is measurable, since h_1 and h_2 are conditional densities. Thus, for all $x_0 \in X_0$, $(h_1 \odot h_2)(x_0)$ is measurable, by Proposition A.2.15. Furthermore, for all $x_0 \in X_0$,

$$\begin{aligned}
&\int_{X_2} (h_1 \odot h_2)(x_0)\, d\nu_2 \\
&= \int_{X_2} \left(\lambda x_2. \int_{X_1} \lambda x_1.h_2(x_0, x_1)(x_2)\; h_1(x_0)\; d\nu_1 \right) d\nu_2 \\
&= \int_{X_1} \left(\lambda x_1. \int_{X_2} \lambda x_2.h_2(x_0, x_1)(x_2)\; h_1(x_0)(x_1)\; d\nu_2 \right) d\nu_1 \qquad \text{[Proposition A.2.16]} \\
&= \int_{X_1} \left(\lambda x_1.h_1(x_0)(x_1) \int_{X_2} \lambda x_2.h_2(x_0, x_1)(x_2)\; d\nu_2 \right) d\nu_1 \\
&= \int_{X_1} h_1(x_0)\, d\nu_1 \\
&= 1.
\end{aligned}$$

Hence, for all $x_0 \in X_0$, $(h_1 \odot h_2)(x_0)$ is a density. Thus $h_1 \odot h_2$ is well-defined.

Also, by Proposition A.2.15, $\lambda(x_0, x_2). \int_{X_1} \lambda x_1.h_2(x_0, x_1)(x_2)\; h_1(x_0)\; d\nu_1 : X_0 \times X_2 \to \mathbb{R}$ is measurable. In other words, $\lambda(x_0, x_2).(h_1 \odot h_2)(x_0)(x_2) : X_0 \times X_2 \to \mathbb{R}$ is measurable. Hence $h_1 \odot h_2 : X_0 \to \mathcal{D}(X_2)$ is a conditional density. □

Example A.3.8. Let $h_1 : X \to \mathcal{D}(\mathbb{N}_0)$ be defined by $h_1(x)(n) = e^{-\lambda(x)} \frac{\lambda(x)^n}{n!}$, for all $x \in X$ and $n \in \mathbb{N}_0$, where $\lambda : X \to \mathbb{R}$ is measurable and $\lambda(x) > 0$, for all $x \in X$. In applications, λ would typically be a simple function, in which case h_1 would be piecewise-constant. Then h_1 is a conditional Poisson density. Let $h_2 : X \times \mathbb{N}_0 \to \mathcal{D}(\mathbb{N}_0)$ be defined by $h_2(x, n)(m) = e^{-\delta(x)} \frac{\delta(x)^m}{m!}$, for all $x \in X$, $n \in \mathbb{N}_0$, and $m \in \mathbb{N}_0$, where $\delta : X \to \mathbb{R}$ is measurable and $\delta(x) > 0$, for all $x \in X$. Then h_2 is also a conditional Poisson density.

Now consider $h_1 \odot h_2 : X \to \mathcal{D}(\mathbb{N}_0)$. Then

$$\begin{aligned}
&(h_1 \odot h_2)(x)(m) \\
&= \int_{\mathbb{N}_0} \lambda n.h_2(x, n)(m)\; h_1(x)\, dc \\
&= \sum_{n \in \mathbb{N}_0} h_2(x, n)(m)\; h_1(x)(n) \\
&= \sum_{n \in \mathbb{N}_0} e^{-\delta(x)} \frac{\delta(x)^m}{m!}\; e^{-\lambda(x)} \frac{\lambda(x)^n}{n!} \\
&= e^{-\delta(x)} \frac{\delta(x)^m}{m!}\; e^{-\lambda(x)} \sum_{n \in \mathbb{N}_0} \frac{\lambda(x)^n}{n!} \\
&= e^{-\delta(x)} \frac{\delta(x)^m}{m!},
\end{aligned}$$

for all $x \in X$ and $m \in \mathbb{N}_0$. Thus $h_1 \odot h_2$ is a conditional Poisson density.

Extensions of Definition A.3.7 to contexts where some of the domain arguments are missing will also be useful. For example, let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$, and $(X_3, \mathcal{A}_3)$ be measurable spaces, $(X_2, \mathcal{A}_2, \nu_2)$ and $(X_4, \mathcal{A}_4, \nu_4)$ σ-finite measure spaces, and $h_1 : X_0 \times X_1 \to$

$\mathcal{D}(X_2)$ and $h_2 : X_0 \times X_3 \times X_1 \times X_2 \to \mathcal{D}(X_4)$ conditional densities. (So domain argument X_3 is missing from h_1.) Define $h_1' : X_0 \times X_3 \times X_1 \to \mathcal{D}(X_2)$ by $h_1'(x_0, x_3, x_1) = h_1(x_0, x_1)$, for all $x_0 \in X_0$, $x_3 \in X_3$, and $x_1 \in X_1$. Then $h_1 \odot h_2 : X_0 \times X_3 \times X_1 \to \mathcal{D}(X_4)$ is defined to be $h_1' \odot h_2$. Thus

$$(h_1 \odot h_2)(x_0, x_3, x_1) = \lambda x_4. \int_{X_2} \lambda x_2. h_2(x_0, x_3, x_1, x_2)(x_4)\ h_1(x_0, x_1)\ d\nu_2,$$

for all $x_0 \in X_0$, $x_3 \in X_3$, and $x_1 \in X_1$.

For probability kernels induced by conditional densities, there is the following connection between fusion of the probability kernels and fusion of the corresponding conditional densities.

Proposition A.3.8. *Let $(X_0, \mathcal{A}_0)$ be a measurable space, $(X_1, \mathcal{A}_1, \nu_1)$ and $(X_2, \mathcal{A}_2, \nu_2)$ σ-finite measure spaces, and $h_1 : X_0 \to \mathcal{D}(X_1)$ and $h_2 : X_0 \times X_1 \to \mathcal{D}(X_2)$ conditional densities. Then*

$$(h_1 \cdot \nu_1) \odot (h_2 \cdot \nu_2) = (h_1 \odot h_2) \cdot \nu_2.$$

Proof. For all $x_0 \in X_0$ and $A_2 \in \mathcal{A}_2$,

$$\begin{aligned}
&((h_1 \cdot \nu_1) \odot (h_2 \cdot \nu_2))(x_0)(A_2) \\
&= \int_{X_1} \lambda x_1. (h_2 \cdot \nu_2)(x_0, x_1)(A_2)\ d(h_1 \cdot \nu_1)(x_0) \\
&= \int_{X_1} \lambda x_1. (h_2 \cdot \nu_2)(x_0, x_1)(A_2)\ h_1(x_0)\ d\nu_1 && \text{[Proposition A.3.3]} \\
&= \int_{X_1} \lambda x_1. \left(\int_{X_2} \mathbf{1}_{A_2}\, h_2(x_0, x_1)\ d\nu_2 \right)\ h_1(x_0)\ d\nu_1 \\
&= \int_{X_1} \left(\lambda x_1. \int_{X_2} \lambda x_2. \mathbf{1}_{A_2}(x_2)\, h_2(x_0, x_1)(x_2)\ h_1(x_0)(x_1)\ d\nu_2 \right) d\nu_1 \\
&= \int_{X_2} \left(\lambda x_2. \int_{X_1} \lambda x_1. \mathbf{1}_{A_2}(x_2)\, h_2(x_0, x_1)(x_2)\ h_1(x_0)(x_1)\ d\nu_1 \right) d\nu_2 && \text{[Proposition A.2.16]} \\
&= \int_{X_2} \mathbf{1}_{A_2} \left(\lambda x_2. \int_{X_1} \lambda x_1. h_2(x_0, x_1)(x_2)\ h_1(x_0)\ d\nu_1 \right) d\nu_2 \\
&= \int_{X_2} \mathbf{1}_{A_2}\, (h_1 \odot h_2)(x_0)\ d\nu_2 \\
&= ((h_1 \odot h_2) \cdot \nu_2)(x_0)(A_2).
\end{aligned}$$

□

The concept of a projective product for conditional densities will be needed.

Definition A.3.8. Let $(X, \mathcal{A})$ be a measurable space, $(Y, \mathcal{B}, \nu)$ a measure space, $f : X \to Y \to \mathbb{R}$ a non-negative function such that $\lambda(x, y).f(x)(y) : X \times Y \to \mathbb{R}$ is measurable, and $h : X \to \mathcal{D}(Y)$ a conditional density such that $0 < \int_Y f(x)\ d(h \cdot \nu)(x) < \infty$, for all $x \in X$. Then the *projective product* $f * h : X \to \mathcal{D}(Y)$ *of* f *and* h is defined by

$$(f * h)(x) = \frac{f(x)\, h(x)}{\int_Y f(x)\, h(x)\ d\nu},$$

for all $x \in X$.

Taking X to be a singleton set in Definition A.3.8, the following special case is obtained. Let $(Y, \mathcal{B}, \nu)$ be a measure space, $f : Y \to \mathbb{R}$ a non-negative measurable function, and $h : \mathcal{D}(Y)$ a density such that $0 < \int_Y f \; d(h \cdot \nu) < \infty$. Then the projective product $f * h : \mathcal{D}(Y)$ of f and h is defined by

$$f * h = \frac{f\, h}{\int_Y f\, h\, d\nu}.$$

Proposition A.3.9. *Let $(X, \mathcal{A})$ be a measurable space, $(Y, \mathcal{B}, \nu)$ a measure space, $f : X \to Y \to \mathbb{R}$ a non-negative function such that $\lambda(x, y).f(x)(y) : X \times Y \to \mathbb{R}$ is measurable, and $h : X \to \mathcal{D}(Y)$ a conditional density such that $0 < \int_Y f(x)\; d(h \cdot \nu)(x) < \infty$, for all $x \in X$. Then $f * h : X \to \mathcal{D}(Y)$ is a conditional density.*

Proof. For all $x \in X$, $(f * h)(x)$ is measurable, using Proposition A.2.6, and $\int_Y (f * h)(x)\; d\nu = 1$. Hence, for all $x \in X$, $(f * h)(x) \in \mathcal{D}(Y)$. Also the mapping $(x, y) \mapsto (f * h)(x)(y)$ is measurable. Hence $f * h : X \to \mathcal{D}(Y)$ is a conditional density. □

Here is a useful property of projective products of conditional densities.

Proposition A.3.10. *Let $(X, \mathcal{A})$ be a measurable space, $(Y, \mathcal{B}, \nu)$ a measure space, and $f : X \to \mathcal{D}(Y)$ and $g : X \to \mathcal{D}(Y)$ conditional densities. Then $(f * g) \cdot \nu = f * (g \cdot \nu)$.*

Proof. For all $x \in X$ and $B \in \mathcal{B}$,

$$\begin{aligned}
&((f * g) \cdot \nu)(x)(B) \\
&= \int_Y \mathbf{1}_B\, (f * g)(x)\; d\nu \\
&= \int_Y \mathbf{1}_B\, \frac{f(x)g(x)}{\int_Y f(x)g(x)\; d\nu}\; d\nu \\
&= \frac{\int_Y \mathbf{1}_B f(x)\; d(g \cdot \nu)(x)}{\int_Y f(x) d(g \cdot \nu)(x)} && \text{[Proposition A.3.6]} \\
&= (f * (g \cdot \nu))(x)(B).
\end{aligned}$$

□

A.4 Topological Properties

To ensure the existence of certain mathematical objects, for example, regular conditional distributions, measurable spaces will need to satisfy some topological conditions. Background material on topology can be found, for example, in [43, 44].

Definition A.4.1. A topological space is *Polish* if it is separable and admits a metric under which it is complete.

Polish spaces are the main ingredient of descriptive set theory, which is the study of the definable subsets of such spaces. ('Definable' subsets are defined explicitly from open sets using set-theoretic operations such as complementation, countable union, and projection, and certainly not the axiom of choice.) This theory has many beautiful connections with logic, topology, functional analysis, probability theory, and other parts of mathematics. The connection with probability theory will be important for the following development of the theory of empirical beliefs. An account of descriptive set theory is given in [85].

The product of countably many Polish spaces is Polish. The sum of countably many Polish spaces is Polish. A set of the form $\bigcap_{n\in\mathbb{N}} U_n$, where each U_n is an open set, is a G_δ set and a set of the form $\bigcup_{n\in\mathbb{N}} F_n$, where each F_n is a closed set, is a F_σ set. Thus every open subset is a G_δ set. Every closed subset of a metric space is a G_δ set [85, Proposition 3.7]. A subspace of a Polish space is Polish iff it is a G_δ set [85, Theorem 3.11].

Here are some typical Polish spaces.

Example A.4.1. $\mathbb{R}$, $\mathbb{C}$, $\mathbb{R}^n$, $\mathbb{C}^n$, $\mathbb{R}^{\mathbb{N}}$, and $\mathbb{C}^{\mathbb{N}}$, with their usual topologies, are Polish. Being closed, the unit interval $\mathbb{I} = [0,1]$, and the unit circle $\mathbb{T} = \{x \in \mathbb{C} : |x| = 1\}$ are Polish.

Being an open subset of $\mathbb{R}$, the open interval $(0,1)$ with its usual topology is Polish. (Note the sublety here: $(0,1)$ is not complete with its usual metric, but there is a metric on $(0,1)$ compatible with its usual topology under which it is complete, and hence it is a Polish space.)

Every separable complete metric space is Polish. Every compact metric space is Polish, being separable and complete.

Every countable set with the discrete topology is Polish.

The space $X^{\mathbb{N}}$, where X is countable and has the discrete topology, is Polish.

Separable Banach spaces are Polish spaces.

Here are some spaces that are not Polish.

Example A.4.2. The rationals $\mathbb{Q}$ with the usual topology is not Polish because it is not a G_δ set in $\mathbb{R}$. (If $\mathbb{Q}$ is supposed to be a G_δ set, a contradiction with a consequence of the Baire category theorem is easily obtained.)

An uncountable product of Polish spaces, each of which consists of at least two points, is not Polish because it is not metrizable (in fact, it is not even first countable).

Definition A.4.2. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. An *isomorphism* between $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ is a function $f : X \to Y$ such that f is a bijection, f is measurable, and $f^{-1} : Y \to X$ is measurable. In this case, $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are said to be *isomorphic*.

Definition A.4.3. A measurable space is a *standard Borel space* if it is isomorphic to a measurable space of the form $(Y, \mathcal{B}(Y))$, for some Polish space Y.

Recall that $\mathcal{B}(Y)$ is the Borel σ-algebra on Y. Note that a standard Borel space is not (necessarily) a topological space, but it does 'inherit' its σ-algebra from a topological space. Furthermore, by using the isomorphism to transfer the topology from Y, a measurable space is a standard Borel space if and only if it has the form $(X, \mathcal{B}(X))$, where X is a Polish space. The class of standard Borel spaces is a large class of measurable spaces that includes those that are likely to be met in the practical application of the theory of empirical beliefs developed here.

The product of countably many standard Borel spaces is standard Borel. The sum of countably many standard Borel spaces is standard Borel. If $(X, \mathcal{A})$ is a standard Borel space and $Y \in \mathcal{A}$, then $(Y, \mathcal{A}|_Y)$ is a standard Borel space [85, Corollary 13.4]. Every singleton subset of a standard Borel space is measurable.

Here is a proof that the product of countably many standard Borel spaces is standard Borel, which provides some insight into the interplay between topological and measurable concepts in the theory of standard Borel spaces.

Proposition A.4.1. *Let I be a countable index set and $(X_i, \mathcal{A}_i)_{i\in I}$ a family of standard Borel spaces. Then $(\prod_{i\in I} X_i, \bigotimes_{i\in I} \mathcal{A}_i)$ is a standard Borel space.*

Proof. For each $i \in I$, since X_i is a standard Borel space, there is a Polish space Y_i such that $(X_i, \mathcal{A}_i)$ is isomorphic to $(Y_i, \mathcal{B}(Y_i))$. Then $\prod_{i\in I} Y_i$, with the product topology, is a Polish space.

Next it is shown that $\bigotimes_{i\in I} \mathcal{B}(Y_i) \subseteq \mathcal{B}(\prod_{i\in I} Y_i)$. By definition, $\bigotimes_{i\in I} \mathcal{B}(Y_i)$ is the smallest σ-algebra on $\prod_i Y_i$ such that, for each $i \in I$, the canonical projection $\pi_i : \prod_{i\in I} Y_i \to Y_i$ is measurable (where Y_i has the σ-algebra $\mathcal{B}(Y_i)$). Thus it is enough to show that each π_i is measurable when $\prod_{i\in I} Y_i$ has the σ-algebra $\mathcal{B}(\prod_{i\in I} Y_i)$. Now the product topology on $\prod_{i\in I} Y_i$ is the smallest topology on $\prod_i Y_i$ such that, for each $i \in I$, the canonical projection $\pi_i : \prod_{i\in I} Y_i \to Y_i$ is continuous. Thus, for all $i \in I$, if O_i is an open subset of Y_i, then $\pi_i^{-1}(O_i)$ is an open subset of $\prod_{i\in I} Y_i$ and therefore a measurable set in $\mathcal{B}(\prod_{i\in I} Y_i)$. Since $\mathcal{B}(Y_i)$ is generated by the topology on Y_i, it is easy to show from this that if D_i is a measurable subset of Y_i, then $\pi_i^{-1}(D_i)$ is a measurable set in $\mathcal{B}(\prod_{i\in I} Y_i)$. Thus each π_i measurable.

Next, it is shown that $\mathcal{B}(\prod_{i\in I} Y_i) \subseteq \bigotimes_{i\in I} \mathcal{B}(Y_i)$. Each Y_i is a separable metrizable space and therefore second countable, that is, there is a countable base $\mathcal{U}_i$ for its topology. Thus each open set in $\prod_{i\in I} Y_i$ is a countable union of sets in $\{\pi_i^{-1}(U_i) \mid U_i \in \mathcal{U}_i, i \in I\} \subseteq \bigotimes_{i\in I} \mathcal{B}(Y_i)$. Therefore each open set in $\prod_{i\in I} Y_i$ is in $\bigotimes_{i\in I} \mathcal{B}(Y_i)$. Hence $\mathcal{B}(\prod_{i\in I} Y_i) \subseteq \bigotimes_{i\in I} \mathcal{B}(Y_i)$.

Finally, $(\prod_{i\in I} X_i, \bigotimes_{i\in I} \mathcal{A}_i)$ is isomorphic to $(\prod_{i\in I} Y_i, \bigotimes_{i\in I} \mathcal{B}(Y_i)$, since $(X_i, \mathcal{A}_i)$ is isomorphic to $(Y_i, \mathcal{B}(Y_i))$, for all $i \in I$. But $\bigotimes_{i\in I} \mathcal{B}(Y_i) = \mathcal{B}(\prod_{i\in I} Y_i)$. Hence the result. □

As an example that is prominent throughout, if X is a standard Borel space, then Proposition A.4.1 shows that $X^{\mathbb{N}_0}$ is also a standard Borel space. The set of constant sequences in $X^{\mathbb{N}_0}$ will play a role in the key concept of constant-valued a.s.; the next result shows that this set, and certain related subsets, are measurable.

Proposition A.4.2. *Let $(X, \mathcal{A})$ be a standard Borel space and, for all $n \in \mathbb{N}_0$, $\alpha_n : X \to X$ a measurable function. Then*

$$\{f \in X^{\mathbb{N}_0} \mid f(n) = \alpha_n(f(n+1)), \text{ for all } n \in \mathbb{N}_0\}$$

is a measurable subset of $X^{\mathbb{N}_0}$.

Proof. Without loss of generality, it can be assumed that X is a Polish space and $\mathcal{A} = \mathcal{B}(X)$. Let $C \triangleq \{f \in X^{\mathbb{N}_0} \mid f(n) = \alpha_n(f(n+1)), \text{ for all } n \in \mathbb{N}_0\}$ and, for all $n \in \mathbb{N}_0$, $C_n \triangleq \{f \in X^{\mathbb{N}_0} \mid f(n) = \alpha_n(f(n+1))\}$, so that $C = \bigcap_{n\in\mathbb{N}_0} C_n$.

By Proposition A.1.4, for all $n \in \mathbb{N}_0$, $\{(x, \alpha_n(x)) \mid x \in X\}$ is a measurable subset of $X \times X$. Consequently, for all $n \in \mathbb{N}_0$, C_n is a measurable subset of $X^{\mathbb{N}_0}$, so that C is a measurable subset of $X^{\mathbb{N}_0}$. □

Here is a key result about standard Borel spaces that will be needed.

Proposition A.4.3. *Every standard Borel space is isomorphic to a standard Borel space of the form $(B, \mathcal{B}(B))$, for some Borel subset B of $\mathbb{R}$.*

This is a consequence of what is known as the Borel isomorphism theorem ([85, Theorem 15.6] or [43, Theorem 13.1.1]): Any two standard Borel spaces are isomorphic iff they have the same cardinality, which moreover is either countable or c $(= 2^{\aleph_0})$.

A technical result due to Carathéodory about joint measurability will be useful.

Proposition A.4.4. *Let $(X, \mathcal{X})$ be a measurable space, $(Y, \mathcal{Y})$ and $(Z, \mathcal{Z})$ topological spaces, and $f : X \times Y \to Z$ a function. Suppose that Y is a separable metrizable space and Z a metrizable space. Suppose also that, for all $y \in Y$, $\lambda x.f(x, y) : X \to Z$ is measurable and for all $x \in X$, $\lambda y.f(x, y) : Y \to Z$ is continuous. Then f is measurable.*

Proof. Note that Y and Z are measurable spaces with their respective Borel σ-algebras. See [4, Lemma 4.51]. □

The next concept is relevant for a useful factorization result.

Definition A.4.4. Let $f : X \to Y$ and $g : X \to Z$ be functions. Then f is *g-measurable* if $\sigma(f) \subseteq \sigma(g)$.

Proposition A.4.5. *Let X be a set, $(Y, \mathcal{Y})$ and $(Z, \mathcal{Z})$ measurable spaces, where Y is standard Borel, and $f : X \to Y$ and $g : X \to Z$ functions. Then f is g-measurable if and only if there exists a measurable function $h : Z \to Y$ such that $f = h \circ g$.*

Proof. See [83, Lemma 1.13]. □

A.5 Random Variables

The fundamental idea of probability theory is that a random experiment is modelled by a distinguished probability space, denoted here by $(\Omega, \mathfrak{S}, \mathsf{P})$, called the *basic probability space*. The set Ω is called the *sample space*. Each $\omega \in \Omega$ is called an *outcome* (that is, a possible outcome of some random experiment). Each $A \in \mathfrak{S}$ is called an *event*. The probability of each event occurring (that is, the probability of the outcome of the experiment being a member of the event) is given by the probability measure P. Generally, events are not directly observed; instead there are functions (called random variables, defined below) from Ω to some measurable space (often $\mathbb{R}$) and the values of the random variables are observed. Indirectly, each such observation gives information about the corresponding event that has occurred. The main mathematical objects of study are the distributions of the random variables.

Definition A.5.1. A *random variable* is a measurable function $f : \Omega \to Y$, where $(\Omega, \mathfrak{S}, \mathsf{P})$ is a probability space and $(Y, \mathcal{B})$ a measurable space. It is said that f is a random variable *in* Y.

The probability measure $\mathsf{P} \circ f^{-1}$ on $\mathcal{B}$ is called the *law* (or *distribution*) of f, and is denoted by $\mathcal{L}(f)$.

For certain cases of Y, terminology other than 'random variable' in Definition A.5.1 is commonly used. In the literature, 'random variable' is often reserved for the case when $Y = \mathbb{R}$, but here 'random variable' will indicate *any* measurable function defined on the basic probability space. If $Y = \mathbb{R}^n$, for some $n > 0$, a random variable is usually called a *random vector*.

If Y is a function space X^T, for some measurable space X and set T, then a random variable will be called a *stochastic process*. (The terminology *random process* is often used in the literature.) In this case, T is usually interpreted as time, and $T = \mathbb{N}_0$ or $T = \mathbb{N}$ in the discrete case, or $T \subseteq \mathbb{R}$ in the continuous case. If $f : \Omega \to X^T$ is a stochastic process, then $f = (f_t)_{t\in T}$, where $f_t : \Omega \to X$ is defined by $f_t = \pi_t \circ f$, for all $t \in T$, and $\pi_t : X^T \to X$ is an evaluation map, for all $t \in T$. Then, by Proposition A.1.5, $f : \Omega \to X^T$ is measurable iff $f_t : \Omega \to X$ is measurable, for all $t \in T$. Thus one can identify the stochastic process f with the indexed family of random variables $(f_t)_{t\in T}$. For each $\omega \in \Omega$, $(f_t(\omega))_{t\in T}$ is called a *path* for the stochastic process.

The distribution of a stochastic process is determined by the set of its finite-dimensional distributions.

Proposition A.5.1. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, X a measurable space, and T a set. Suppose that $f : \Omega \to X^T$ and $g : \Omega \to X^T$ are stochastic processes. Then $\mathcal{L}(f) = \mathcal{L}(g)$ iff $\mathcal{L}((f_{t_1}, \ldots, f_{t_n})) = \mathcal{L}((g_{t_1}, \ldots, g_{t_n}))$, for each finite subset $\{t_1, \ldots, t_n\} \subseteq T$.*

Proof. See [83, Proposition 3.2]. □

Note. To capture the intuitive idea of a *process*, the index set T needs to be infinite. Also a random variable taking values in an infinite product space is a stochastic process. To see this, consider an indexed family $(X_t)_{t\in T}$ of measurable spaces such that $\bigcup_{t\in T} X_t$ is measurable. Then

$$\prod_{t\in T} X_t = \{f \mid f \in (\bigcup_{t\in T} X_t)^T, \text{ where } f(t) \in X_t, \text{ for all } t \in T\},$$

so that $\prod_{t\in T} X_t \subseteq (\bigcup_{t\in T} X_t)^T$. The case of most interest in this book is when $T = \mathbb{N}$, so the values of the corresponding stochastic processes are sequences.

Definition A.5.2. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $f : \Omega \to \mathbb{R}$ a random variable that is integrable. Then the *expectation* $\mathsf{E}_\mathsf{P}(f)$ of f is defined by

$$\mathsf{E}_\mathsf{P}(f) = \int_\Omega f \, d\mathsf{P}.$$

Definition A.5.3. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $f : \Omega \to \mathbb{R}$ a random variable that is square integrable. Then the *variance* $\mathsf{Var}_\mathsf{P}(f)$ of f is defined by

$$\mathsf{Var}_\mathsf{P}(f) = \int_\Omega f^2 \, d\mathsf{P} - \left(\int_\Omega f \, d\mathsf{P}\right)^2.$$

If the probability measure is clear from the context, $\mathsf{E}_\mathsf{P}(f)$ is abbreviated to $\mathsf{E}(f)$ and $\mathsf{Var}_\mathsf{P}(f)$ is abbreviated to $\mathsf{Var}(f)$. Note that $\mathsf{Var}(f) = \mathsf{E}(f^2) - \mathsf{E}(f)^2 = \mathsf{E}(f - \mathsf{E}(f))^2$. Also $\sqrt{(\mathsf{Var}(f)}$ is called the *standard deviation* of f.

Definition A.5.4. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $f_i : \Omega \to \mathbb{R}$ a random variable that is square integrable, for $i = 1, \ldots, n$. Then the *covariance* $\mathsf{Cov}_{\mathsf{P}}(f_1, \ldots, f_n)$ of $f_1, \ldots, f_n$ is defined by

$$\mathsf{Cov}_{\mathsf{P}}(f_1, \ldots, f_n) = \mathsf{E}(\prod_{i=1}^{n}(f_i - \mathsf{E}(f_i))).$$

If the probability measure is clear from the context, $\mathsf{Cov}_{\mathsf{P}}(f_1, \ldots, f_n)$ is abbreviated to $\mathsf{Cov}(f_1, \ldots, f_n)$. Note that $\mathsf{Cov}(f_1, \ldots, f_n) = \mathsf{E}(\prod_{i=1}^{n} f_i) - \prod_{i=1}^{n} \mathsf{E}(f_i)$.

Definition A.5.5. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $\mathcal{E}$ and $\mathcal{F}$ sub-σ-algebras of $\mathfrak{S}$. Then $\mathcal{E}$ and $\mathcal{F}$ are *independent*, denoted $\mathcal{E} \perp\!\!\!\perp \mathcal{F}$, if

$$\mathsf{P}(E \cap F) = \mathsf{P}(E)\,\mathsf{P}(F),$$

for all $E \in \mathcal{E}$ and $F \in \mathcal{F}$. More generally, an indexed family $(\mathcal{E}_i)_{i \in I}$ of sub-σ-algebras of $\mathfrak{S}$ is *independent* if

$$\mathsf{P}(\bigcap_{j \in J} E_j) = \prod_{j \in J} \mathsf{P}(E_j),$$

for all finite subsets J of I and all $E_j \in \mathcal{E}_j$.

Definition A.5.6. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(X_i, \mathcal{A}_i)$ measurable spaces, for $i \in I$. An indexed family $(f_i : \Omega \to X_i)_{i \in I}$ of random variables is *independent* if $(\sigma(f_i))_{i \in I}$ is independent.

In other words, $(f_i : \Omega \to X_i)_{i \in I}$ is independent if

$$\mathsf{P}(\bigcap_{j \in J} f_j^{-1}(A_j)) = \prod_{j \in J} \mathsf{P}(f_J^{-1}(A_j)),$$

for all finite subsets J of I and all $A_j \in \mathcal{A}_j$.

Definition A.5.7. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(X, \mathcal{A})$ a measurable space. A sequence $(f_n : \Omega \to X)_{n \in \mathbb{N}}$ of random variables is *independent and identically distributed* (i.i.d.) if $(f_n : \Omega \to X)_{n \in \mathbb{N}}$ is independent and each f_n has the same distribution.

Later, in Example A.8.1, it will be shown that it is possible to construct a sequence $(f_n)_{n \in \mathbb{N}}$ of independent random variables of given distributions. This construction provides a typical setting that concerns a sequence of i.i.d. random variables. For this setting, the basic probability space is $(\Omega, \mathfrak{S}, \mathsf{P})$, where $\Omega = \prod_{n \in \mathbb{N}} \Omega_n$, $\mathfrak{S} = \bigotimes_{n \in \mathbb{N}} \mathfrak{S}_n$, and $\mathsf{P} = \bigotimes_{n \in \mathbb{N}} \mathsf{P}_n$, for some sequence $(\Omega_n, \mathfrak{S}_n, \mathsf{P}_n)_{n \in \mathbb{N}}$ of probability spaces. Thus each outcome in Ω is a sequence $(\omega_n)_{n \in \mathbb{N}}$, where each $\omega_n \in \Omega_n$. The sequence $(f_n : \Omega \to X_n)_{n \in \mathbb{N}}$ of random variables has the property that each f_n factors through Ω_n. Thus, for all $n \in \mathbb{N}$, there exists a measurable function $g_n : \Omega_n \to X_n$ such that $f_n = g_n \circ \pi_n$, where $\pi_n : \Omega \to \Omega_n$ is the canonical projection. The argument of Example A.8.1 shows that the sequence $(f_n)_{n \in \mathbb{N}}$ is independent. The crucial fact needed to establish this is that each f_n factors through Ω_n.

Now imagine the process of generating a sequence of random i.i.d. values for this setting. First, an outcome $\omega = (\omega_n)_{n\in\mathbb{N}} \in \Omega$ is sampled according to P. Each component ω_n of this sequence provides the basis for a value in the sequence of random values. For example, in a coin tossing setting, ω could be an infinite sequence of heads or tails. For each n, $f_n(\omega) = g_n(\omega_n)$ and this provides the nth element of the sequence of random values. The random values are independent because, in effect, each f_n uses only the nth component of ω.

Here is a key property that events in a probability space may have.

Definition A.5.8. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space. If $\mathsf{P}(A) = 1$, for some $A \in \mathfrak{S}$, then A is said to happen *almost surely*.

This is denoted by P-a.s., if the relevant probability measure is P, or, by a.s., if the relevant probability measure is clear from the context.

Use will be made of the strong law of large numbers for the theory of particle filters.

Proposition A.5.2. *(Strong law of large numbers) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(\xi_n : \Omega \to \mathbb{R})_{n\in\mathbb{N}}$ a sequence of i.i.d. random variables that are integrable. Let $M \triangleq \mathsf{E}(\xi_1)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \xi_j = M \text{ a.s.}$$

Proof. There are many variants of this result; see, for example, [43, Theorem 8.3.5], [81, Theorem 20.1], [83, Theorem 4.23], or [87, Theorem 5.17]. □

More explicitly, the conclusion of Proposition A.5.2 is that

$$\mathsf{P}(\{\omega \in \Omega \mid \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \xi_j(\omega) = M\}) = 1$$

or, in other words,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \xi_j = \lambda\omega.M \text{ a.s.}$$

The strong law of large numbers provides the theoretical basis for Monte Carlo integration. Suppose that we want to evaluate an integral of the form

$$\int_X fh \, d\nu,$$

where $(X, \mathcal{A}, \nu)$ is a measure space, $h : \mathcal{D}(X)$ is a density on $(X, \mathcal{A}, \nu)$, and $f : X \to \mathbb{R}$ is $(h \cdot \nu)$-integrable. In effect, by Proposition A.3.3, the above integral is equal to the integral

$$\int_X f \, d(h \cdot \nu),$$

of f with respect to the probability measure $h \cdot \nu$. Here is the basic result about Monte Carlo integration.

Proposition A.5.3. *(Monte Carlo integration) Let* $(\Omega, \mathfrak{S}, \mathsf{P})$ *be a probability space,* $(X, \mathcal{A})$ *a measurable space,* $(\eta_n : \Omega \to X)_{n\in\mathbb{N}}$ *an i.i.d. sequence of random variables with distribution* μ*, and* $f : X \to \mathbb{R}$ *a* μ*-integrable function. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (f \circ \eta_j) = \int_X f \, d\mu \text{ a.s.}$$

Proof. It is easy to show that $(f \circ \eta_n : \Omega \to \mathbb{R})_{n\in\mathbb{N}}$ is an i.i.d. sequence of integrable random variables. Since the distribution of η_n is μ, it follows that $\mathsf{P} \circ \eta_n^{-1} = \mu$, for all $n \in \mathbb{N}$. Also, for all $n \in \mathbb{N}$, $\mathsf{E}(f \circ \eta_n) = \int_\Omega f \circ \eta_n \, d\mathsf{P} = \int_X f \, d(\mathsf{P} \circ \eta_n^{-1}) = \int_X f \, d\mu$. The result now follows directly from the Strong Law of Large Numbers. □

Thus, for a suitably large N, one can use $\frac{1}{N} \sum_{j=1}^{N} f(\eta_j(\omega))$ to approximate $\int_X f \, d\mu$, for almost all choices of $\omega \in \Omega$.

Proposition A.5.3 is commonly used in the situation where there is a measure ν on $(X, \mathcal{A})$ and a density $h : X \to \mathbb{R}$ on $(X, \mathcal{A}, \nu)$ such that $\mu = h \cdot \nu$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (f \circ \eta_j) = \int_X fh \, d\nu \text{ a.s.}$$

Next it is shown that a special case of Proposition A.5.3 gives an empirical approximation of a probability measure.

Notation. For a random variable $\eta : \Omega \to X$, the notation δ_η means $\lambda\omega.\delta_{\eta(\omega)}$.

Thus $\delta_\eta : \Omega \to \Delta(X)$ is a random measure.

Definition A.5.9. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ a measurable space, and $\eta_j : \Omega \to X$ a random variable, for $j = 1, \ldots, n$. Then the random measure

$$\frac{1}{n} \sum_{j=1}^{n} \delta_{\eta_j}$$

is called the *empirical measure* (*determined by* $\eta_1, \ldots, \eta_n$).

For all $\omega \in \Omega$, $\frac{1}{n} \sum_{j=1}^{n} \delta_{\eta_j(\omega)}$ is a Dirac mixture measure.

Proposition A.5.4. *(Approximation by empirical measures) Let* $(\Omega, \mathfrak{S}, \mathsf{P})$ *be a probability space,* $(X, \mathcal{A})$ *a measurable space,* μ *a probability measure on* $(X, \mathcal{A})$*, and* $(\eta_n : \Omega \to X)_{n\in\mathbb{N}}$ *an i.i.d. sequence of random variables with distribution* μ*. Then, for all* $A \in \mathcal{A}$*,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \delta_{\eta_j}(A) = \mu(A) \text{ a.s.}$$

Proof. Letting $f \triangleq \mathbf{1}_A$ in Proposition A.5.3, the result follows immediately. □

Thus, for a suitably large N, one can use $\frac{1}{N}\sum_{j=1}^{N}\delta_{\eta_j(\omega)}$ to approximate μ, for almost all choices of $\omega \in \Omega$.

Propositions A.5.2, A.5.3, and A.5.4 show that Monte Carlo methods essentially depend on being able to sample from a given distribution. Given a basic probability space $(\Omega, \mathfrak{S}, \mathsf{P})$, a measure space $(X, \mathcal{A})$, and a random variable $\eta : \Omega \to X$, the task is to sample a random value from the distribution $\mathcal{L}(\eta)$ on X. This corresponds to the physical interpretation of the basic probability space as the space of outcomes from a random experiment and a random variable as providing some property of interest of an outcome. Everything flows from the existence of random outcomes $\omega \in \Omega$ generated by the distribution P. However, in practice, the basic probability space is usually not available and instead one has to work directly with (postulated) explicit distributions for the various random variables. Thus, for a random variable $\eta : \Omega \to X$, it is necessary to directly sample a value $x \in X$ from the distribution $\mu \triangleq \mathcal{L}(\eta)$, denoted $x \sim \mu$. (The value x is equal to $\eta(\omega)$, for some $\omega \in \Omega$, where ω is sampled from the distribution P.) Starting from an i.i.d. sequence of uniform random values on $(0, 1)$, one can develop sampling methods for common distributions that are needed in applications. See, for example, [14, Ch. 11], [113, Ch. 23], and [136].

Now, as an illustration, consider how sampling is used in applications of Proposition A.5.3. For a sufficiently large N, let $(x_j)_{j=1}^{N}$ be a family of independent random values in X sampled from the distribution μ. Precisely, for some $\omega \in \Omega$, $x_j = \eta_j(\omega)$, for $j = 1, \ldots, N$, where $(\eta_j : \Omega \to X)_{j=1}^{N}$ is a family of i.i.d. random variables with distribution μ. Then, by Proposition A.5.3,

$$\int_X f \, d\mu \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j).$$

This kind of approximation of integrals will be used in particle filters in Section 4.3.

The concept of conditional expectation, which is a random variable rather than a single value, will be heavily used.

Definition A.5.10. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $\mathcal{F}$ a sub-σ-algebra of $\mathfrak{S}$, and $f : \Omega \to \mathbb{R}$ a random variable. Then a *conditional expectation* of f with respect to $\mathcal{F}$ is an $\mathcal{F}$-measurable random variable $g : \Omega \to \mathbb{R}$ such that $\mathsf{E}(\mathbf{1}_B g) = \mathsf{E}(\mathbf{1}_B f)$, for all $B \in \mathcal{F}$.

Conditional expectation is denoted by $\mathsf{E}_{\mathsf{P}}(f \,|\, \mathcal{F})$. When the probability measure P is clear from the context, it is denoted more simply by $\mathsf{E}(f \,|\, \mathcal{F})$. Thus

$$\int_\Omega \mathbf{1}_B \mathsf{E}(f \,|\, \mathcal{F}) \, d\mathsf{P} = \int_\Omega \mathbf{1}_B f \, d\mathsf{P},$$

for all $B \in \mathcal{F}$. A common case is where $\mathcal{F}$ is the σ-algebra generated by some measurable function $g : \Omega \to Y$, for some Y, in which case $\mathsf{E}(f \,|\, \sigma(g))$ is denoted by $\mathsf{E}(f \,|\, g)$.

Here are some remarks that provide some intuition about the concept of conditional expectation. First, σ-algebras are used to model information in the following way. Given some σ-algebra $\mathcal{F} \subseteq \mathfrak{S}$, suppose a random experiment is performed and an outcome $\omega \in \Omega$ determined, but the value of ω is not revealed. Instead, for each set in the σ-algebra $\mathcal{F}$, one is told whether or not ω is in the set. The larger the σ-algebra, the more information it provides. At the other extreme, the trivial σ-algebra $\{\{\}, \Omega\}$ provides no information.

Suppose now that f is a real-valued random variable that is measurable with respect to $\mathcal{F}$. Then, even though the information in $\mathcal{F}$ may not be enough to determine ω, it *is* enough to determine the value $f(\omega)$. (To see this, consider the measurable sets $f^{-1}(\{r\})$, for all $r \in \mathbb{R}$.) In this sense, f is *determined by* $\mathcal{F}$.

Note that, strictly, it is not the σ-algebra that contains the information; the actual information is whether or not the outcome belongs to any particular set in the σ-algebra. However, with the understanding that the knowledge of which measurable sets contain the outcome and which do not *is taken for granted*, one can speak of a σ-algebra modelling information in this way.

Now consider conditional expectation itself. The intuition is that $\mathsf{E}(f \,|\, \mathcal{F})$ is an $\mathcal{F}$-measurable random variable that provides the 'best' estimate of the value of f given the information in $\mathcal{F}$. By the remarks just above, since $\mathsf{E}(f \,|\, \mathcal{F})$ is $\mathcal{F}$-measurable, it can be evaluated using only the information in $\mathcal{F}$, that is, $\mathsf{E}(f \,|\, \mathcal{F})$ is determined by $\mathcal{F}$. Also the meaning of 'best' is given by the requirement that $\mathsf{E}(f \,|\, \mathcal{F})$ and f should have the same average over each set in $\mathcal{F}$, that is, $\mathsf{E}(\mathbf{1}_B \,\mathsf{E}(f \,|\, \mathcal{F})) = \mathsf{E}(\mathbf{1}_B f)$, for all $B \in \mathcal{F}$. If f is $\mathcal{F}$-measurable, then f itself provides such an estimate. At the other extreme, if $\sigma(f)$ and $\mathcal{F}$ are independent, then $\mathcal{F}$ provides no useful information and the best estimate available is the expectation of f. (See Proposition A.5.6, Part 8.) The more interesting situation is when f lies between these two extremes, in which case the information in $\mathcal{F}$ is relevant to the evaluation of the values for f, but is not sufficient to determine them exactly. For this situation, the general definition of $\mathsf{E}(f \,|\, \mathcal{F})$ is needed.

The next result establishes the existence of conditional expectations.

Proposition A.5.5. *For any probability space $(\Omega, \mathfrak{S}, \mathsf{P})$, sub-$\sigma$-algebra $\mathcal{F}$ of $\mathfrak{S}$, and integrable random variable $f : \Omega \to \mathbb{R}$, a conditional expectation $\mathsf{E}(f \,|\, \mathcal{F})$ exists, and any two such conditional expectations are equal a.s.*

Proof. See [43, p.337]. The proof uses the Radon-Nikodym theorem (Proposition A.2.12). □

Conditional probability is a special case of conditional expectation.

Definition A.5.11. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $\mathcal{F}$ a sub-σ-algebra of $\mathfrak{S}$, and $A \in \mathfrak{S}$. Then the *conditional probability* of A with respect to $\mathcal{F}$, denoted by $\mathsf{P}(A \,|\, \mathcal{F})$, is defined to be $\mathsf{E}(\mathbf{1}_A \,|\, \mathcal{F})$.

In addition, $\mathsf{P}(A \,|\, \sigma(g))$ is denoted by $\mathsf{P}(A \,|\, g)$.

Example A.5.1. This example relates the concept of conditional expectation just introduced to a more elementary notion of this concept. Let $f : \Omega \to \mathbb{R}$ be a random variable and $B \in \mathfrak{S}$ a measurable set such that $\mathsf{P}(B) > 0$. Define the (elementary) conditional expectation $\mathsf{E}(f \,|\, B)$ of f with respect to B by

$$\mathsf{E}(f \,|\, B) = \frac{\mathsf{E}(\mathbf{1}_B f)}{\mathsf{P}(B)}.$$

Note that $\mathsf{E}(f \,|\, B)$ is a number. In fact, $\mathsf{E}(f \,|\, B)$ is the expectation of f with respect to the (elementary) conditional probability measure $\lambda A.\mathsf{P}(A \,|\, B)$ defined by

$$\mathsf{P}(A \,|\, B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)},$$

for all $A \in \mathfrak{S}$.

Now consider a countable partition $(B_i)_{i\in I}$ of Ω such that $\mathsf{P}(B_i) > 0$, for all $i \in I$. Let $\mathcal{F}$ be the σ-algebra generated by $(B_i)_{i\in I}$. Thus $\mathcal{F}$ is the set of all sets of the form $\bigcup_{j\in J} B_j$, for all $J \subseteq I$. Let $f : \Omega \to \mathbb{R}$ be a random variable. Then

$$\mathsf{E}(f \,|\, \mathcal{F})(\omega) = \mathsf{E}(f \,|\, B_i), \text{ if } \omega \in B_i,$$

for almost all $\omega \in \Omega$. Thus $\mathsf{E}(f \,|\, \mathcal{F})$ is almost surely constant on each B_i. It is now shown that the definition for $\mathsf{E}(f \,|\, \mathcal{F})$ is correct. Let $g : \Omega \to \mathbb{R}$ be defined by $g(\omega) = \mathsf{E}(f \,|\, B_i)$, if $\omega \in B_i$, for all $\omega \in \Omega$. Then g is clearly $\mathcal{F}$-measurable. Next consider $B \triangleq \bigcup_{j\in J} B_j \in \mathcal{F}$ and let $j \in J$. Then $\mathsf{E}(\mathbf{1}_{B_j} g) = \int_\Omega \mathbf{1}_{B_j} g \; d\mathsf{P} = \int_\Omega \mathbf{1}_{B_j} \mathsf{E}(f \,|\, B_j) \; d\mathsf{P} = \int_\Omega \mathbf{1}_{B_j} \frac{\mathsf{E}(\mathbf{1}_{B_j} f)}{\mathsf{P}(B_j)} \; d\mathsf{P} = \mathsf{E}(\mathbf{1}_{B_j} f)$. It follows that $\mathsf{E}(\mathbf{1}_B g) = \mathsf{E}(\mathbf{1}_B f)$, for all $B \in \mathcal{F}$. Thus g is a conditional expectation of f with respect to $\mathcal{F}$.

Here is a collection of useful properties of conditional expectations.

Proposition A.5.6. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $\mathcal{F}$ a sub-σ-algebra of $\mathfrak{S}$, and $f : \Omega \to \mathbb{R}$ and $g : \Omega \to \mathbb{R}$ integrable functions. Then the following hold.*

1. $\mathsf{E}(\alpha f + \beta g \mid \mathcal{F}) = \alpha \mathsf{E}(f \mid \mathcal{F}) + \beta \mathsf{E}(g \mid \mathcal{F})$ *a.s., for all* $\alpha, \beta \in \mathbb{R}$.
2. $f \geq 0$ *implies* $\mathsf{E}(f \mid \mathcal{F}) \geq 0$ *a.s.*
3. $\mathsf{E}(|\mathsf{E}(f \mid \mathcal{F})|) \leq \mathsf{E}(|f|)$.
4. $0 \leq f_n \uparrow f$ *implies* $\mathsf{E}(f_n \mid \mathcal{F}) \uparrow \mathsf{E}(f \mid \mathcal{F})$ *a.s.*
5. $\mathsf{E}(fg \mid \mathcal{F}) = f\mathsf{E}(g \mid \mathcal{F})$ *a.s., whenever f is $\mathcal{F}$-measurable.*
6. $\mathsf{E}(f\, \mathsf{E}(g \mid \mathcal{F})) = \mathsf{E}(g\, \mathsf{E}(f \mid \mathcal{F})) = \mathsf{E}(\mathsf{E}(f \mid \mathcal{F})\, \mathsf{E}(g \mid \mathcal{F}))$.
7. $\mathsf{E}(\mathsf{E}(f \mid \mathcal{G}) \mid \mathcal{F}) = \mathsf{E}(f \mid \mathcal{F})$ *a.s., whenever* $\mathcal{F} \subseteq \mathcal{G}$.
8. *If $\sigma(f)$ and $\mathcal{F}$ are independent, then* $\mathsf{E}(f \mid \mathcal{F}) = \lambda\omega.\mathsf{E}(f)$ *a.s.*
9. *If* $\mathsf{P}(A) \in \{0, 1\}$, *for all* $A \in \mathcal{F}$, *then* $\mathsf{E}(f \mid \mathcal{F}) = \lambda\omega.\mathsf{E}(f)$ *a.s.*

Proof. See [83, Theorem 6.1] and [87, Theorem 8.14]. □

Note that, by Part 2 of Proposition A.5.6, $0 \leq \mathsf{P}(A \mid \mathfrak{F}) \leq 1$ a.s., for all $A \in \mathfrak{S}$.

Here is another useful property of conditional expectations.

Proposition A.5.7. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ sub-σ-algebras of $\mathfrak{S}$ such that $\mathcal{F} \subseteq \mathcal{G} \subseteq \mathcal{H}$. Let $f : \Omega \to \mathbb{R}$ be an integrable random variable such that $\mathsf{E}(f \,|\, \mathcal{F}) = \mathsf{E}(f \,|\, \mathcal{H})$ a.s. Then $\mathsf{E}(f \,|\, \mathcal{F}) = \mathsf{E}(f \,|\, \mathcal{G})$ a.s.*

Proof. By definition, $\mathsf{E}(f \,|\, \mathcal{H}) : \Omega \to \mathbb{R}$ is $\mathcal{H}$-measurable and $\int_\Omega \mathbf{1}_B \mathsf{E}(f \,|\, \mathcal{H}) \; d\mathsf{P} = \int_\Omega \mathbf{1}_B f \; d\mathsf{P}$, for all $B \in \mathcal{H}$. Hence $\int_\Omega \mathbf{1}_B \mathsf{E}(f \,|\, \mathcal{F}) \; d\mathsf{P} = \int_\Omega \mathbf{1}_B f \; d\mathsf{P}$, for all $B \in \mathcal{G}$, and $\mathsf{E}(f \,|\, \mathcal{F})$ is $\mathcal{G}$-measurable. Hence $\mathsf{E}(f \,|\, \mathcal{F}) = \mathsf{E}(f \,|\, \mathcal{G})$ a.s. □

Proposition A.5.8. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $\mathcal{F}$ a sub-σ-algebra of $\mathfrak{S}$, $(f_n)_{n\in\mathbb{N}}$ a sequence of integrable functions, where each $f_n : \Omega \to \mathbb{R}$, and $f, g : \Omega \to \mathbb{R}$ integrable functions. Suppose that $\lim_{n\to\infty} f_n = f$ a.s. and $|f_n| \leq g$, for all $n \in \mathbb{N}$. Then*

$$\lim_{n\to\infty} \mathsf{E}(f_n \mid \mathcal{F}) = \mathsf{E}(f \mid \mathcal{F}) \text{ a.s.}$$

Proof. See [87, Theorem 8.14, Part (viii)]. □

The next result shows the effect of changing the probability measure.

Proposition A.5.9. *(Change of measure) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $\mathcal{F}$ a sub-σ-algebra of $\mathfrak{S}$.*

1. *Let $g : \Omega \to \mathbb{R}$ be a density and $h : \Omega \to \mathbb{R}$ a random variable. Then*

$$\mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F}) = \frac{\mathsf{E}_{\mathsf{P}}(gh \mid \mathcal{F})}{\mathsf{E}_{\mathsf{P}}(g \mid \mathcal{F})} \quad g \cdot \mathsf{P}\text{-a.s.}$$

2. *Let $g : \Omega \to \mathbb{R}$ and $h : \Omega \to \mathbb{R}$ be densities. Then*

$$\mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F}) = \frac{\mathsf{E}_{h\cdot\mathsf{P}}(g \mid \mathcal{F})\, \mathsf{E}_{\mathsf{P}}(h \mid \mathcal{F})}{\mathsf{E}_{\mathsf{P}}(g \mid \mathcal{F})} \quad (g \cdot \mathsf{P} + h \cdot \mathsf{P})/2\text{-a.s.}$$

3. *Let $A, B \in \mathfrak{S}$ satisfy $\mathsf{P}(A) > 0$ and $\mathsf{P}(B) > 0$. Then*

$$\mathsf{P}|_A(B \mid \mathcal{F}) = \frac{\mathsf{P}|_B(A \mid \mathcal{F})\, \mathsf{P}(B \mid \mathcal{F})}{\mathsf{P}(A \mid \mathcal{F})} \quad (\mathsf{P}|_A + \mathsf{P}|_B)/2\text{-a.s.}$$

Proof. 1. Note that $\mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F})\, \mathsf{E}_{\mathsf{P}}(g \mid \mathcal{F})$ is $\mathcal{F}$-measurable. Also, for all $B \in \mathcal{F}$,

$$\begin{aligned}
&\int_\Omega \mathbf{1}_B\, \mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F})\, \mathsf{E}_{\mathsf{P}}(g \mid \mathcal{F})\, d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_B\, \mathsf{E}_{\mathsf{P}}(\mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F})\, g \mid \mathcal{F})\, d\mathsf{P} \qquad [\mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F}) \text{ is } \mathcal{F}\text{-measurable}] \\
&= \int_\Omega \mathbf{1}_B\, \mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F})\, g\, d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_B\, \mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F})\, d(g \cdot \mathsf{P}) \\
&= \int_\Omega \mathbf{1}_B\, h\, d(g \cdot \mathsf{P}) \\
&= \int_\Omega \mathbf{1}_B\, gh\, d\mathsf{P}.
\end{aligned}$$

Hence $\mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F})\, \mathsf{E}_{\mathsf{P}}(g \mid \mathcal{F}) = \mathsf{E}_{\mathsf{P}}(gh \mid \mathcal{F})$ $g \cdot \mathsf{P}$-a.s. The result follows.

2. From Part 1, it follows that, $(g \cdot \mathsf{P} + h \cdot \mathsf{P})/2$-almost surely,

$$\mathsf{E}_{g\cdot\mathsf{P}}(h \mid \mathcal{F}) = \frac{\mathsf{E}_{\mathsf{P}}(gh \mid \mathcal{F})}{\mathsf{E}_{\mathsf{P}}(g \mid \mathcal{F})} = \frac{\mathsf{E}_{h\cdot\mathsf{P}}(g \mid \mathcal{F})\, \mathsf{E}_{\mathsf{P}}(h \mid \mathcal{F})}{\mathsf{E}_{\mathsf{P}}(g \mid \mathcal{F})}.$$

3. Let $g \triangleq \mathbf{1}_A/\mathsf{P}(A)$ and $h = \mathbf{1}_B/\mathsf{P}(B)$. From Part 2, it follows that, $(\mathbf{1}_A/\mathsf{P}(A) \cdot \mathsf{P} + \mathbf{1}_B/\mathsf{P}(B) \cdot \mathsf{P})/2$-almost surely,

$$\mathsf{E}_{\mathbf{1}_A/\mathsf{P}(A)\cdot\mathsf{P}}(\mathbf{1}_B/\mathsf{P}(B) \mid \mathcal{F}) = \frac{\mathsf{E}_{\mathbf{1}_B/\mathsf{P}(B)\cdot\mathsf{P}}(\mathbf{1}_A/\mathsf{P}(A) \mid \mathcal{F})\ \mathsf{E}_{\mathsf{P}}(\mathbf{1}_B/\mathsf{P}(B) \mid \mathcal{F})}{\mathsf{E}_{\mathsf{P}}(\mathbf{1}_A/\mathsf{P}(A) \mid \mathcal{F})}.$$

Hence, $(\mathbf{1}_A/\mathsf{P}(A) \cdot \mathsf{P} + \mathbf{1}_B/\mathsf{P}(B) \cdot \mathsf{P})/2$-almost surely,

$$(\mathbf{1}_A/\mathsf{P}(A) \cdot \mathsf{P})(B \mid \mathcal{F}) = \frac{(\mathbf{1}_B/\mathsf{P}(B) \cdot \mathsf{P})(A \mid \mathcal{F})\ \mathsf{P}(B \mid \mathcal{F})}{\mathsf{P}(A \mid \mathcal{F})}.$$

Now note that $\mathbf{1}_A/\mathsf{P}(A) \cdot \mathsf{P}$ is just $\mathsf{P}|_A$, the normalized restriction of P to A. Similarly, $\mathbf{1}_B/\mathsf{P}(B) \cdot \mathsf{P} = \mathsf{P}|_B$. Hence the result. □

The next two results give a useful decomposition of a conditional expectation in the context of a product space.

Proposition A.5.10. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $(X_i, \mathcal{A}_i)$ a measurable space for $i = 0, 1, 2$. Suppose that $f_0 : \Omega \to X_0$, $f_1 : \Omega \to X_1$, and $f_2 : \Omega \to X_2$ are measurable. Then, for all $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$,*

$$\mathsf{E}(\mathbf{1}_{(f_1,f_2)^{-1}(A_1\times A_2)} \mid f_0) = \mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\ \mathsf{E}(\mathbf{1}_{f_2^{-1}(A_2)} \mid (f_0, f_1)) \mid f_0) \text{ a.s.}$$

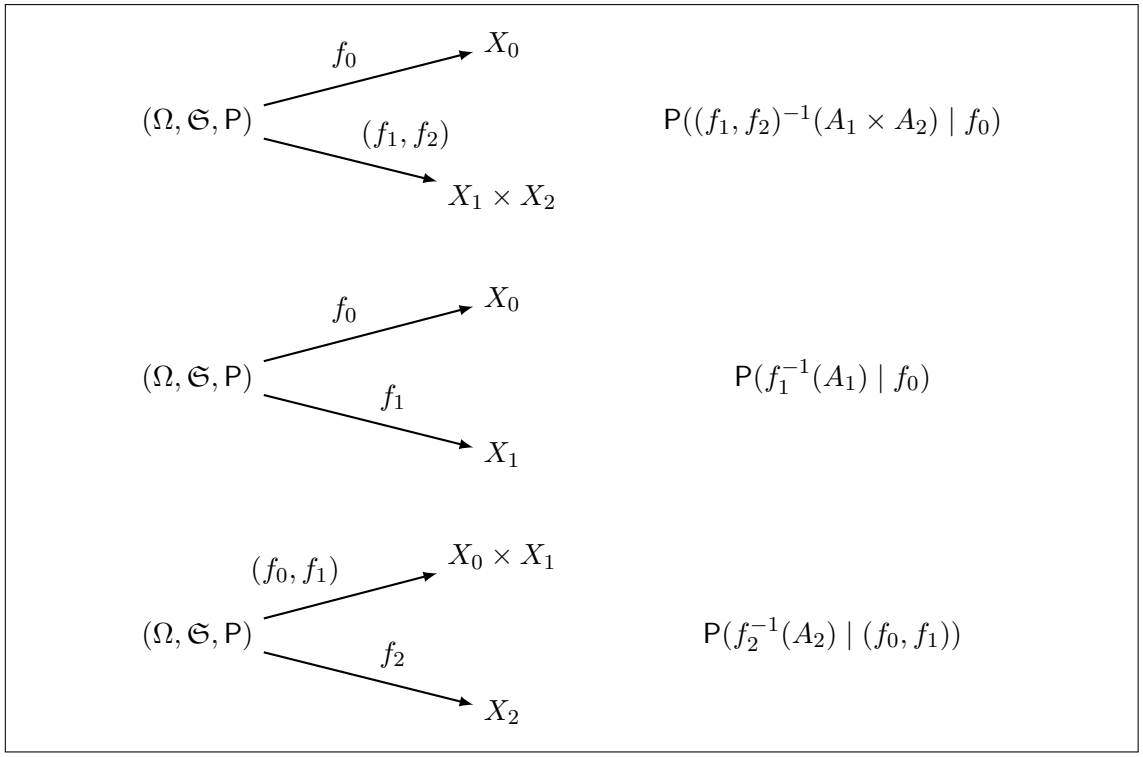

Figure A.1: Setting for Proposition A.5.10

Proof. Suppose that $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$. Then, almost surely,

$$\begin{aligned}
&\mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \mathsf{E}(\mathbf{1}_{f_2^{-1}(A_2)} \mid (f_0, f_1)) \mid f_0) \\
&= \mathsf{E}(\mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \mathbf{1}_{f_2^{-1}(A_2)} \mid (f_0, f_1)) \mid f_0) \\
&\qquad\qquad [\text{Proposition A.5.6, Part 5; } \mathbf{1}_{f_1^{-1}(A_1)} \text{ is } \sigma(f_0, f_1)\text{-measurable}] \\
&= \mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \mathbf{1}_{f_2^{-1}(A_2)} \mid f_0) \qquad [\text{Proposition A.5.6, Part 7; } \sigma(f_0) \subseteq \sigma(f_0, f_1)] \\
&= \mathsf{E}(\mathbf{1}_{(f_1,f_2)^{-1}(A_1 \times A_2)} \mid f_0).
\end{aligned}$$

□

Figure A.1 illustrates that Proposition A.5.10 shows that the conditional probability $\mathsf{P}((f_1, f_2)^{-1}(A_1 \times A_2) \mid f_0)$ is 'factored' into the 'product' of the conditional probabilities $\mathsf{P}(f_1^{-1}(A_1) \mid f_0)$ and $\mathsf{P}(f_2^{-1}(A_2) \mid (f_0, f_1))$.

Several versions of Bayes theorem will be needed. Here is a version for conditional expectations.

Proposition A.5.11. *(Bayes theorem for conditional expectations) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $(X_i, \mathcal{A}_i)$ a measurable space for $i = 0, 1, 2$. Suppose that $f_0 : \Omega \to X_0$, $f_1 : \Omega \to X_1$, and $f_2 : \Omega \to X_2$ are measurable. Then, for all $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$,*

$$\mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \mathsf{E}(\mathbf{1}_{f_2^{-1}(A_2)} \mid (f_0, f_1)) \mid f_0) = \mathsf{E}(\mathbf{1}_{f_2^{-1}(A_2)}\, \mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)} \mid (f_0, f_2)) \mid f_0) \text{ a.s.}$$

Proof. The result follows immediately from Proposition A.5.10. □

Proposition A.5.10 extends, of course, to the general case. (See Figure A.2.)

Proposition A.5.12. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $(X_i, \mathcal{A}_i)$ a measurable space for $i = 0, \ldots, n$. Suppose that $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{i=1}^n X_i$ are measurable, where $f \triangleq (f_1, \ldots, f_n)$. Then, for all $A_1 \in \mathcal{A}_1, \ldots, A_n \in \mathcal{A}_n$,*

$$\mathsf{E}(\mathbf{1}_{f^{-1}(\prod_{i=1}^n A_i)} \mid f_0) = \mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \mathsf{E}(\mathbf{1}_{f_2^{-1}(A_2)} \ldots \mathsf{E}(\mathbf{1}_{f_n^{-1}(A_n)} \mid (f_0, \ldots, f_{n-1}) \ldots) \mid (f_0, f_1)) \mid f_0) \text{ a.s.}$$

Proof. The proof is by induction on n. For $n = 1$, the result is obvious. Suppose, for the inductive step, that the result holds for $n - 1$. Suppose that $A_1 \in \mathcal{A}_1$, ..., $A_n \in \mathcal{A}_n$. Then, almost surely,

$$\begin{aligned}
&\mathsf{E}(\mathbf{1}_{f^{-1}(\prod_{i=1}^n A_i)} \mid f_0) \\
&= \mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \mathsf{E}(\mathbf{1}_{(f_2,\ldots,f_n)^{-1}(\prod_{i=2}^n A_i)} \mid (f_0, f_1)) \mid f_0) \qquad [\text{Proposition A.5.10}] \\
&= \mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \mathsf{E}(\mathbf{1}_{f_2^{-1}(A_2)} \ldots \mathsf{E}(\mathbf{1}_{f_n^{-1}(A_n)} \mid (f_0, \ldots, f_{n-1}) \ldots) \mid (f_0, f_1)) \mid f_0). \\
&\qquad\qquad [\text{Induction hypothesis}]
\end{aligned}$$

This establishes the result. □

Proposition A.5.13. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(\Omega', \mathfrak{S}')$, $(X, \mathcal{A})$, and $(Y, \mathcal{B})$ measurable spaces, $h : \Omega \to \Omega'$ a random variable, and $f : \Omega' \to X$ and $g : \Omega' \to Y$ measurable functions. Then, for all $B \in \mathcal{B}$,*

$$(\mathsf{P} \circ h^{-1})(g^{-1}(B) \,|\, f) \circ h = \mathsf{P}((g \circ h)^{-1}(B) \,|\, f \circ h) \text{ a.s.}$$

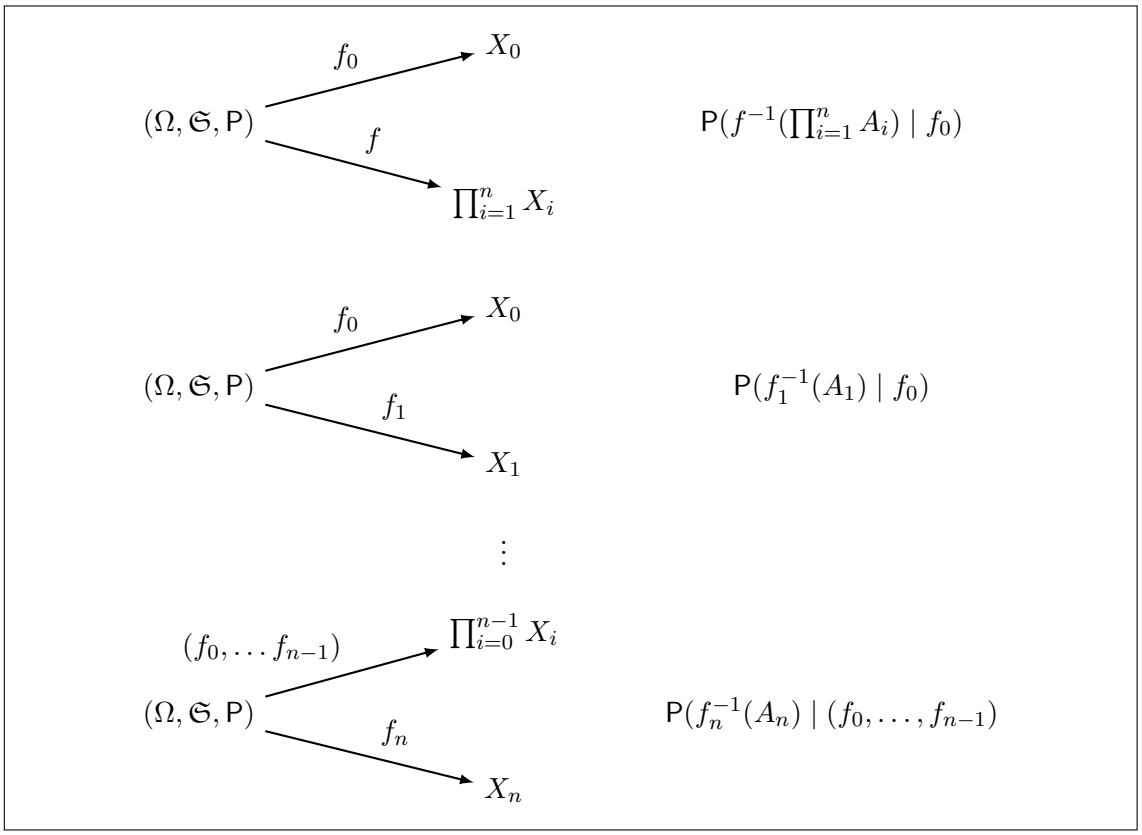

Figure A.2: Setting for Proposition A.5.12

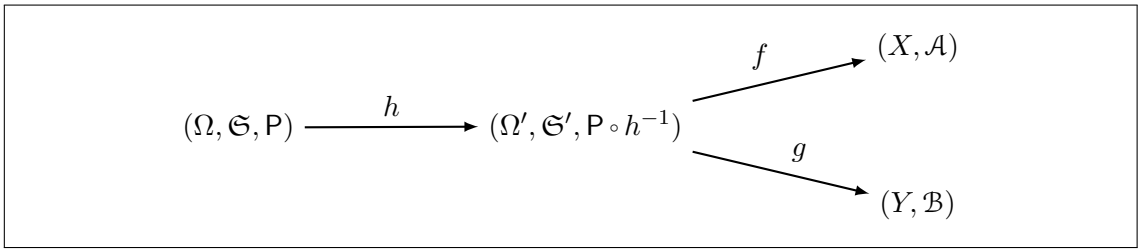

Figure A.3: Setting for Proposition A.5.13

Proof. The conditional probability $\mathsf{P}((g \circ h)^{-1}(B) \,|\, f \circ h)$ is defined a.s. by

$$\int_\Omega \mathbf{1}_C \; \mathsf{P}((g \circ h)^{-1}(B) \,|\, f \circ h) \; d\mathsf{P} = \int_\Omega \mathbf{1}_C \; \mathbf{1}_{(g \circ h)^{-1}(B)} \; d\mathsf{P},$$

for all $C \in \sigma(f \circ h)$. If $C \in \sigma(f \circ h)$, then there exists $A \in \mathcal{A}$ such that $C = h^{-1}(f^{-1}(A))$. Then, for all $B \in \mathcal{B}$,

$$\begin{aligned} &\int_\Omega \mathbf{1}_C \; ((\mathsf{P} \circ h^{-1})(g^{-1}(B) \,|\, f) \circ h) \; d\mathsf{P} \\ = &\int_\Omega (\mathbf{1}_{f^{-1}(A)} \; (\mathsf{P} \circ h^{-1})(g^{-1}(B) \,|\, f)) \circ h \; d\mathsf{P} \end{aligned}$$

$$
\begin{aligned}
&= \int_{\Omega'} \mathbf{1}_{f^{-1}(A)}\ (\mathsf{P}\circ h^{-1})(g^{-1}(B)\,|\,f)\ d(\mathsf{P}\circ h^{-1}) && \text{[Proposition A.2.14]}\\
&= \int_{\Omega'} \mathbf{1}_{f^{-1}(A)}\ \mathbf{1}_{g^{-1}(B)}\ d(\mathsf{P}\circ h^{-1}) && \text{[Definition of cond. probability]}\\
&= \int_{\Omega} (\mathbf{1}_{f^{-1}(A)}\ \mathbf{1}_{g^{-1}(B)})\circ h\ d\mathsf{P} && \text{[Proposition A.2.14]}\\
&= \int_{\Omega} \mathbf{1}_C\ \mathbf{1}_{(g\circ h)^{-1}(B)}\ d\mathsf{P}.
\end{aligned}
$$

Hence, for all $B \in \mathcal{B}$, $(\mathsf{P}\circ h^{-1})(g^{-1}(B)\,|\,f)\circ h = \mathsf{P}((g\circ h)^{-1}(B)\,|\,f\circ h)$ a.s. □

Proposition A.5.13 can be extended to conditional expectations.

Proposition A.5.14. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(\Omega', \mathfrak{S}')$ a measurable space, $h : \Omega \to \Omega'$ a random variable, $k : \Omega' \to \mathbb{R}$ a measurable function, and $\mathcal{F}$ a sub-σ-algebra of $\mathfrak{S}'$. Then*

$$\mathsf{E}_{\mathsf{P}\circ h^{-1}}(k\,|\,\mathcal{F})\circ h = \mathsf{E}_{\mathsf{P}}(k\circ h\,|\,h^{-1}(\mathcal{F}))\ \text{a.s.}$$

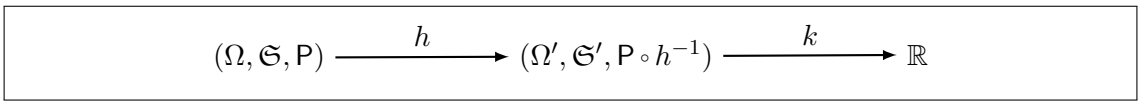

Figure A.4: Setting for Proposition A.5.14

Proof. Since h is measurable and $\mathcal{F} \subseteq \mathfrak{S}'$, $h^{-1}(\mathcal{F}) \subseteq \mathfrak{S}$. The conditional expectation $\mathsf{E}_{\mathsf{P}}(k\circ h\,|\,h^{-1}(\mathcal{F}))$ is defined a.s. by

$$\int_{\Omega} \mathbf{1}_A\ \mathsf{E}_{\mathsf{P}}(k\circ h\,|\,h^{-1}(\mathcal{F}))\ d\mathsf{P} = \int_{\Omega} \mathbf{1}_A\ (k\circ h)\ d\mathsf{P},$$

for all $A \in h^{-1}(\mathcal{F})$. If $A \in h^{-1}(\mathcal{F})$, then there exists $B \in \mathcal{F}$ such that $A = h^{-1}(B)$. Then

$$
\begin{aligned}
&\int_{\Omega} \mathbf{1}_A\ \mathsf{E}_{\mathsf{P}\circ h^{-1}}(k\,|\,\mathcal{F})\circ h\ d\mathsf{P}\\
&= \int_{\Omega} (\mathbf{1}_B\ \mathsf{E}_{\mathsf{P}\circ h^{-1}}(k\,|\,\mathcal{F}))\circ h\ d\mathsf{P}\\
&= \int_{\Omega'} \mathbf{1}_B\ \mathsf{E}_{\mathsf{P}\circ h^{-1}}(k\,|\,\mathcal{F})\ d(\mathsf{P}\circ h^{-1}) && \text{[Proposition A.2.14]}\\
&= \int_{\Omega'} \mathbf{1}_B\ k\ d(\mathsf{P}\circ h^{-1}) && \text{[Definition of conditional expectation]}\\
&= \int_{\Omega} (\mathbf{1}_B\ k)\circ h\ d\mathsf{P} && \text{[Proposition A.2.14]}\\
&= \int_{\Omega} \mathbf{1}_A\ (k\circ h)\ d\mathsf{P}.
\end{aligned}
$$

Hence $\mathsf{E}_{\mathsf{P}\circ h^{-1}}(k\,|\,\mathcal{F})\circ h = \mathsf{E}_{\mathsf{P}}(k\circ h\,|\,h^{-1}(\mathcal{F}))$ a.s. □

Note that Proposition A.5.13 is a special case of Proposition A.5.14. Just let $k \triangleq \mathbf{1}_{g^{-1}(C)}$ and $\mathcal{F} = \sigma(f)$. Then $k \circ h = \mathbf{1}_{h^{-1}(g^{-1}(C))}$ and $h^{-1}(\mathcal{F}) = \sigma(f \circ h)$.

From a practical point of view, conditional probabilities are not convenient to work with; instead, since empirical beliefs, transition models, and observation models are probability kernels, it is more convenient to deal with what is known as regular conditional distributions which are probability kernels that 'represent' conditional probabilities in a certain precise sense. The following proposition is the relevant key property of conditional probabilities, perhaps the most important foundational result needed in the theory of empirical beliefs. The proposition states the sense in which a regular conditional distribution represents a conditional probability. Here the restriction to standard Borel spaces becomes important since the result does not hold if this restriction is dropped. However, as mentioned earlier, all the measurable spaces one is ever likely to meet in practice are standard Borel spaces.

Proposition A.5.15. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(Y, \mathcal{B})$ a standard Borel space, $g : \Omega \to Y$ a random variable, and $\mathcal{F}$ a sub-σ-algebra of $\mathfrak{S}$. Then there exists an $\mathcal{F}$-measurable probability kernel $\nu : \Omega \to \mathcal{P}(Y)$ such that, for all $B \in \mathcal{B}$,*

$$\mathsf{P}(g^{-1}(B) \,|\, \mathcal{F}) = \lambda\omega.\nu(\omega)(B) \text{ a.s.}$$

Proof. See [87, Theorems 8.29 and 8.37]. The idea of the proof is to assume first that Y is the set of the real numbers and exploit the fact that the rationals are a countable dense subset of the reals to construct, using a rather technical argument, a set of distribution functions on Y indexed by $\omega \in \Omega$ and hence a probability kernel $\nu : \Omega \to \mathcal{P}(Y)$. From there, using Proposition A.4.3, it is easy to lift the result to standard Borel spaces. □

In fact, a refinement of Proposition A.5.15 will be needed for which the sub-σ-algebra $\mathcal{F}$ is generated by a random variable $f : \Omega \to X$.

Proposition A.5.16. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ a measurable space, $(Y, \mathcal{B})$ a standard Borel space, and $f : \Omega \to X$ and $g : \Omega \to Y$ random variables. Then there exists a probability kernel $\mu : X \to \mathcal{P}(Y)$ such that, for all $B \in \mathcal{B}$,*

$$\mathsf{P}(g^{-1}(B) \,|\, f) = \lambda\omega.\mu(f(\omega))(B) \text{ a.s.}$$

Furthermore, μ is unique $\mathcal{L}(f)$-a.e.

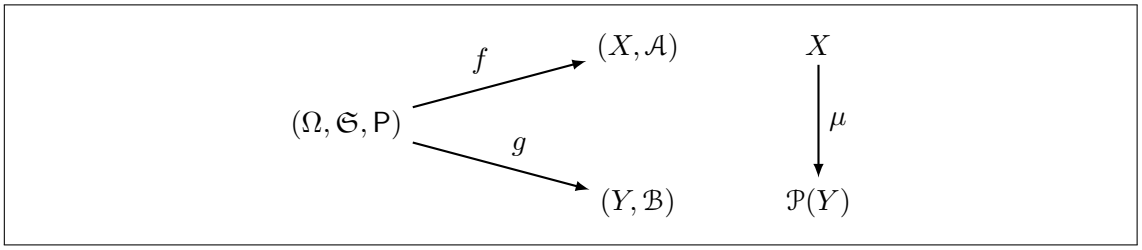

Figure A.5: Setting for Proposition A.5.16

Proof. See [83, Theorem 6.3]. □

Proposition A.5.15 is a special case of Proposition A.5.16: in Proposition A.5.16, just let $(X, \mathcal{A})$ be $(\Omega, \mathcal{F})$ and f the identity mapping from $(\Omega, \mathfrak{S})$ to $(\Omega, \mathcal{F})$. Note that $\sigma(f) = \mathcal{F}$.

Here is the definition of the key concept of a regular conditional distribution.

Definition A.5.12. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ measurable spaces, and $f : \Omega \to X$ and $g : \Omega \to Y$ random variables. A probability kernel $\mu : X \to \mathcal{P}(Y)$ is called a *regular conditional distribution* of g given f if, for all $B \in \mathcal{B}$,

$$\mathsf{P}(g^{-1}(B) \,|\, f) = \lambda\omega.\mu(f(\omega))(B) \text{ a.s.}$$

Proposition A.5.16 shows that regular conditional distributions exist if $(Y, \mathcal{B})$ is a standard Borel space. Note that Proposition A.2.4 shows that, for all $B \in \mathcal{B}$, $\lambda\omega.\mu(f(\omega))(B) : \Omega \to \mathbb{R}$ is $\sigma(f)$-measurable.

Under the conditions of Definition A.5.12, suppose that $\sigma(f)$ and $\sigma(\mathbf{1}_{g^{-1}(B)})$ are independent. This happens, for example, if X is a singleton set, in which case $\sigma(f) = \{\emptyset, \Omega\}$. Then, by Part 8 of Proposition A.5.6, $\mathsf{P}(g^{-1}(B) \,|\, f) = \lambda\omega.\mathsf{E}(\mathbf{1}_{g^{-1}(B)})$ a.s. Now $\mathsf{E}(\mathbf{1}_{g^{-1}(B)}) = \mathsf{P}(g^{-1}(B))$. Hence $\mathsf{P}(g^{-1}(B) \,|\, f) = \lambda\omega.\mathsf{P}(g^{-1}(B))$ a.s. Thus it is possible to identify the conditional probability $\mathsf{P}(g^{-1}(B) \,|\, f)$ with the probability value $\mathsf{P}(g^{-1}(B))$. Furthermore, the probability kernel $\mu : X \to \mathcal{P}(Y)$ defined by $\mu = \lambda x.(\mathsf{P} \circ g^{-1})$ is a regular conditional distribution of g given f.

In the context of Definition A.5.12, if $\mu : X \to \mathcal{P}(Y)$ is a regular conditional distribution of g given f, then Proposition A.5.21 below gives a simple connection between the law of g and the law of f:

$$\mathsf{P} \circ g^{-1} = \lambda B. \int_X \lambda x.\mu(x)(B)\; d(\mathsf{P} \circ f^{-1}).$$

In other words,

$$\mathcal{L}(g) = \mathcal{L}(f) \odot \mu.$$

Also, if $x \in X$, then the conditional probability $\mathsf{P}(g^{-1}(B) \,|\, f)$ is almost surely constant on $f^{-1}(\{x\})$. (If singleton sets in X are measurable, which happens if X is a standard Borel space, for example, then $f^{-1}(\{x\})$ is an event in Ω.)

Here is an example to make the concept of regular conditional distribution more concrete.

Example A.5.2. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\mathfrak{S} = 2^\Omega$, and $\mathsf{P}(\{\omega_1\}) = \frac{1}{2}$, $\mathsf{P}(\{\omega_2\}) = \frac{1}{2}$, and $\mathsf{P}(\{\omega_3\}) = \frac{1}{4}$. Let $X = \{a, b\}$, $\mathcal{A} = 2^X$, and $f : \Omega \to X$ be defined by $f(\omega_1) = a = f(\omega_2)$ and $f(\omega_3) = b$. Let $Y = \{c, d\}$, $\mathcal{B} = 2^Y$, and $g : \Omega \to Y$ be defined by $g(\omega_1) = c = g(\omega_3)$ and $g(\omega_2) = d$. Then f and g are random variables. Note that $\sigma(f) = \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3\}, \Omega\}$.

Now

$$\mathsf{P}(g^{-1}(\{c\}) \,|\, f) = \mathsf{P}(\{\omega_1, \omega_3\} \,|\, f) = \mathsf{E}(\mathbf{1}_{\{\omega_1, \omega_3\}} \,|\, f)$$
$$\mathsf{P}(g^{-1}(\{d\}) \,|\, f) = \mathsf{P}(\{\omega_2\} \,|\, f) = \mathsf{E}(\mathbf{1}_{\{\omega_2\}} \,|\, f).$$

Furthermore, by Example A.5.1,

$$\begin{aligned}
&\mathsf{E}(\mathbf{1}_{\{\omega_1,\omega_3\}} \mid f)(\omega_1) = \mathsf{E}(\mathbf{1}_{\{\omega_1,\omega_3\}} \mid \{\omega_1,\omega_2\}) = \frac{\mathsf{E}(\mathbf{1}_{\{\omega_1,\omega_2\}}\mathbf{1}_{\{\omega_1,\omega_3\}})}{\mathsf{P}(\{\omega_1,\omega_2\})} = \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}\\
&\mathsf{E}(\mathbf{1}_{\{\omega_1,\omega_3\}} \mid f)(\omega_2) = \mathsf{E}(\mathbf{1}_{\{\omega_1,\omega_3\}} \mid \{\omega_1,\omega_2\}) = \frac{2}{3}\\
&\mathsf{E}(\mathbf{1}_{\{\omega_1,\omega_3\}} \mid f)(\omega_3) = \mathsf{E}(\mathbf{1}_{\{\omega_1,\omega_3\}} \mid \{\omega_3\}) = \frac{\mathsf{E}(\mathbf{1}_{\{\omega_3\}}\mathbf{1}_{\{\omega_1,\omega_3\}})}{\mathsf{P}(\{\omega_3\})} = 1\\
&\mathsf{E}(\mathbf{1}_{\{\omega_2\}} \mid f)(\omega_1) = \mathsf{E}(\mathbf{1}_{\{\omega_2\}} \mid \{\omega_1,\omega_2\}) = \frac{\mathsf{E}(\mathbf{1}_{\{\omega_1,\omega_2\}}\mathbf{1}_{\{\omega_2\}})}{\mathsf{P}(\{\omega_1,\omega_2\})} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}\\
&\mathsf{E}(\mathbf{1}_{\{\omega_2\}} \mid f)(\omega_2) = \mathsf{E}(\mathbf{1}_{\{\omega_2\}} \mid \{\omega_1,\omega_2\}) = \frac{1}{3}\\
&\mathsf{E}(\mathbf{1}_{\{\omega_2\}} \mid f)(\omega_3) = \mathsf{E}(\mathbf{1}_{\{\omega_2\}} \mid \{\omega_3\}) = \frac{\mathsf{E}(\mathbf{1}_{\{\omega_3\}}\mathbf{1}_{\{\omega_2\}})}{\mathsf{P}(\{\omega_3\})} = 0.
\end{aligned}$$

Now define the probability kernel $\mu : X \to \mathcal{P}(Y)$ by $\mu(a)(\{c\}) = \frac{2}{3}$, $\mu(a)(\{d\}) = \frac{1}{3}$, $\mu(b)(\{c\}) = 1$, and $\mu(b)(\{d\}) = 0$. Hence

$$\begin{aligned}
&\mathsf{P}(g^{-1}(\{c\}) \mid f)(\omega_1) = \mathsf{P}(\{\omega_1,\omega_3\} \mid f)(\omega_1) = \frac{2}{3} = \mu(a)(\{c\}) = \mu(f(\omega_1)(\{c\})\\
&\mathsf{P}(g^{-1}(\{c\}) \mid f)(\omega_2) = \mathsf{P}(\{\omega_1,\omega_3\} \mid f)(\omega_2) = \frac{2}{3} = \mu(a)(\{c\}) = \mu(f(\omega_2)(\{c\})\\
&\mathsf{P}(g^{-1}(\{c\}) \mid f)(\omega_3) = \mathsf{P}(\{\omega_1,\omega_3\} \mid f)(\omega_3) = 1 = \mu(b)(\{c\}) = \mu(f(\omega_3)(\{c\})\\
&\mathsf{P}(g^{-1}(\{d\}) \mid f)(\omega_1) = \mathsf{P}(\{\omega_2\} \mid f)(\omega_1) = \frac{1}{3} = \mu(a)(\{d\}) = \mu(f(\omega_1)(\{d\})\\
&\mathsf{P}(g^{-1}(\{d\}) \mid f)(\omega_2) = \mathsf{P}(\{\omega_2\} \mid f)(\omega_2) = \frac{1}{3} = \mu(a)(\{d\}) = \mu(f(\omega_2)(\{d\})\\
&\mathsf{P}(g^{-1}(\{d\}) \mid f)(\omega_3) = \mathsf{P}(\{\omega_2\} \mid f)(\omega_3) = 0 = \mu(b)(\{d\}) = \mu(f(\omega_3)(\{d\}).
\end{aligned}$$

It follows that $\mathsf{P}(g^{-1}(B) \mid f) = \lambda\omega.\mu(f(\omega))(B)$, for all $B \in \mathcal{B}$. Hence μ is a regular conditional distribution of g given f.

Proposition A.5.17. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(\Omega', \mathfrak{S}')$ and $(X, \mathcal{A})$ measurable spaces, $(Y, \mathcal{B})$ a standard Borel space, $h : \Omega \to \Omega'$ a random variable, and $f : \Omega' \to X$ and $g : \Omega' \to Y$ measurable functions. Then there exists a probability kernel $\eta : X \to \mathcal{P}(Y)$ such that $\mathsf{P}((g \circ h)^{-1}(B) \mid f \circ h) = \lambda\omega.\eta((f \circ h)(\omega))(B)$ a.s., for all $B \in \mathcal{B}$, and a probability kernel $\xi : X \to \mathcal{P}(Y)$ such that $(\mathsf{P} \circ h^{-1})(g^{-1}(B) \mid f) = \lambda\omega'.\xi(f(\omega')(B)$ $(\mathsf{P} \circ h^{-1})$-a.s., for all $B \in \mathcal{B}$. Furthermore, $\eta = \xi$ $\mathcal{L}(f \circ h)$-a.e.*

Proof. The probability kernels η and ξ exist, by Proposition A.5.16. Suppose that $B \in \mathcal{B}$. Then, almost surely,

$$\begin{aligned}
&\lambda\omega.\eta(f(h(\omega))(B)\\
&= \mathsf{P}((g \circ h)^{-1}(B) \mid f \circ h)\\
&= \lambda\omega.(\mathsf{P} \circ h^{-1})(g^{-1}(B) \mid f)(h(\omega)) && \text{[Proposition A.5.13]}\\
&= \lambda\omega.\xi(f(h(\omega))(B).
\end{aligned}$$

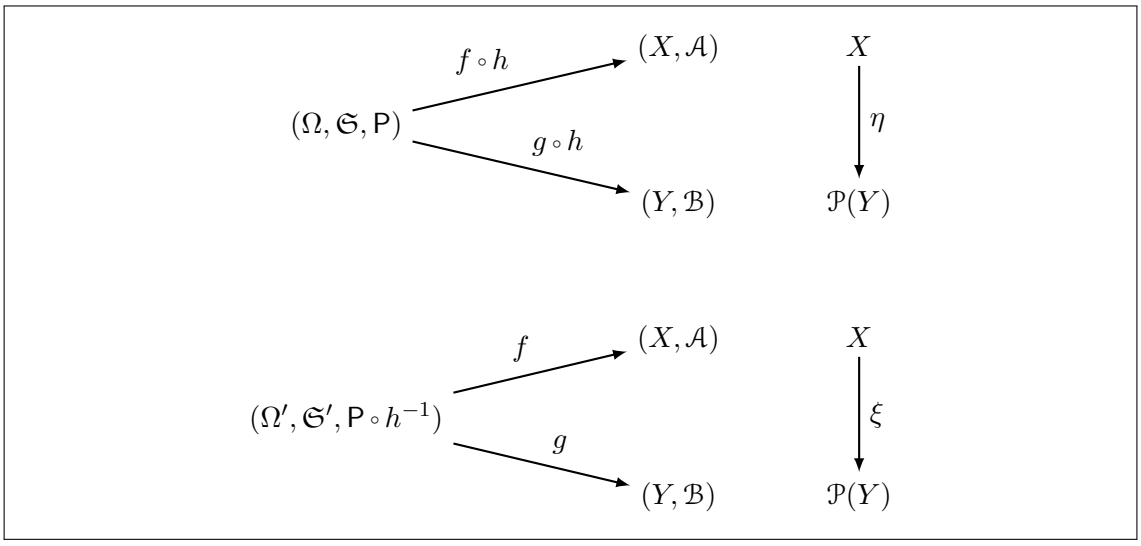

Figure A.6: Setting for Proposition A.5.17

Thus $\eta(f(h(\omega)) = \xi(f(h(\omega))$, for almost all $\omega \in \Omega$. Since

$$h^{-1}(f^{-1}(\{x \in X \mid \eta(x) \neq \xi(x)\})) = \{\omega \in \Omega \mid \eta(f(h(\omega))) \neq \xi(f(h(\omega)))\},$$

it follows that

$$\mathsf{P}(h^{-1}(f^{-1}(\{x \in X \mid \eta(x) \neq \xi(x)\}))) = \mathsf{P}(\{\omega \in \Omega \mid \eta(f(h(\omega))) \neq \xi(f(h(\omega)))\}) = 0,$$

and so $\eta = \xi \quad \mathcal{L}(f \circ h)$-a.e. □

Example A.5.3. A common task with Bayesian networks is to compute the posterior distribution of a set of query variables, given values for a set of evidence variables [88, Section 2.5.1], [140, Section 14.4]. Here is the setting for this.

Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i)$ a product of measurable spaces, and $h : \Omega \to \prod_{i=1}^n X_i$ a random variable. Let $Q \subset \{1, \ldots, n\}$ be the set of indices of the query variables and $E \subset \{1, \ldots, n\}$ be the set of indices of the evidence variables, where $Q \cap E = \emptyset$. Finally, let $\pi_Q : \prod_{i=1}^n X_i \to \prod_{i \in Q} X_i$ and $\pi_E : \prod_{i=1}^n X_i \to \prod_{i \in E} X_i$ be the respective projections.

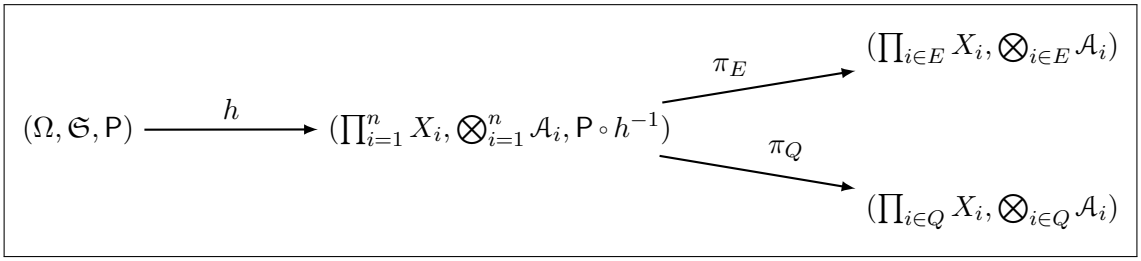

Figure A.7: Setting for Example A.5.3

Then the required task is, for some particular values of the evidence variables (that determine a corresponding event), to compute the conditional probability

$$\mathsf{P}((\pi_Q \circ h)^{-1}(C) \mid \pi_E \circ h),$$

for all $C \in \bigotimes_{i \in Q} \mathcal{A}_i$. By Proposition A.5.13, this conditional probability can be written a.s. in the form

$$(\mathsf{P} \circ h^{-1})(\pi_Q^{-1}(C) \mid \pi_E) \circ h.$$

Furthermore, under the assumption that $(\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i)$ is a standard Borel space, Proposition A.5.17 shows that there exists a probability kernel $\eta : \prod_{i \in E} X_i \to \mathcal{P}(\prod_{i \in Q} X_i)$ such that

$$\mathsf{P}((\pi_Q \circ h)^{-1}(C) \,|\, \pi_E \circ h) = \lambda\omega.\eta((\pi_E \circ h)(\omega))(C) \text{ a.s.},$$

for all $C \in \bigotimes_{i \in Q} \mathcal{A}_i$. Alternatively, there exists a probability kernel $\xi : \prod_{i \in E} X_i \to \mathcal{P}(\prod_{i \in Q} X_i)$ such that

$$(\mathsf{P} \circ h^{-1})(\pi_Q^{-1}(C) \mid \pi_E) = \lambda\omega'.\eta(\pi_E(\omega'))(C) \quad (\mathsf{P} \circ h^{-1})\text{-a.s.},$$

for all $C \in \bigotimes_{i \in Q} \mathcal{A}_i$.

The following result gives a useful property of regular conditional distributions.

Proposition A.5.18. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ measurable spaces, and $f : \Omega \to X$ and $g : \Omega \to Y$ random variables. Suppose there is a probability kernel $\mu : X \to \mathcal{P}(Y)$ such that, for all $B \in \mathcal{B}$, $\mathsf{P}(g^{-1}(B) \,|\, f) = \lambda\omega.\mu(f(\omega))(B)$ a.s. Let $h : X \times Y \to \mathbb{R}$ be a measurable function such that $\mathsf{E}(|h \circ (f, g)|) < \infty$. Then*

$$\mathsf{E}(h \circ (f, g) \,|\, f) = \lambda\omega. \int_Y \lambda y.h(f(\omega), y)\, d\mu(f(\omega)) \text{ a.s.}$$

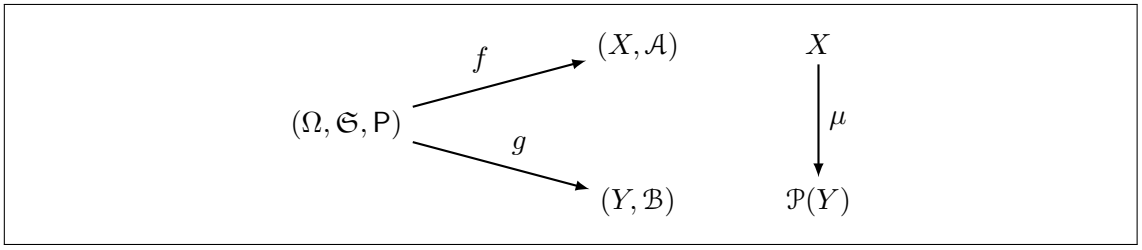

Figure A.8: Setting for Proposition A.5.18

Proof. See [83, Theorem 6.4] (or [87, Theorem 8.38] for the special case given by Equation A.5.1 below). □

A special case of Proposition A.5.18 is of particular interest. Let $h : Y \to \mathbb{R}$ be a measurable function such that $\mathsf{E}(|h \circ g|) < \infty$. Define $h' : X \times Y \to \mathbb{R}$ by $h'(x, y) = h(y)$, for all $x \in X$ and $y \in Y$. Then, according to Proposition A.5.18,

$$\mathsf{E}(h' \circ (f, g) \mid f) = \lambda\omega. \int_Y \lambda y.h'(f(\omega), y)\ d\mu(f(\omega)) \text{ a.s.}$$

That is,

$$\mathsf{E}(h \circ g \mid f) = \lambda\omega. \int_Y h\ d\mu(f(\omega)) \text{ a.s.} \tag{A.5.1}$$

This means that, for all $x \in X$, computing the integral

$$\int_Y h\ d\mu(x)$$

is the same as computing the (constant) value of the conditional expectation $\mathsf{E}(h \circ g \mid f)$ on the event $f^{-1}(\{x\})$. In the proofs of Propositions A.7.10 and A.7.12 below, it is the special case of Proposition A.5.18 given by Equation A.5.1 that is used.

By the way, let $B \in \mathcal{B}$ and h be $\mathbf{1}_B : Y \to \mathbb{R}$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(g^{-1}(B) \mid f) \\
&= \mathsf{E}(\mathbf{1}_{g^{-1}(B)} \mid f) \\
&= \mathsf{E}(\mathbf{1}_B \circ g \mid f) \\
&= \lambda\omega. \int_Y \mathbf{1}_B\ d\mu(f(\omega)) && \text{[Equation A.5.1]} \\
&= \lambda\omega.\mu(f(\omega))(B).
\end{aligned}$$

This shows that the equation $\mathsf{P}(g^{-1}(B) \mid f) = \lambda\omega.\mu(f(\omega))(B)$ a.s. of Proposition A.5.16 is a special case of Equation A.5.1.

In the even more special case where μ is a probability measure (rather than a probability kernel), then an expectation (rather than a conditional expectation) is computed. Let $g : \Omega \to Y$ be a random variable and $\mu = \mathsf{P} \circ g^{-1}$. Then, by Proposition A.2.14,

$$\mathsf{E}(h \circ g) = \int_Y h\ d\mu.$$

Here are two results relevant for calculating conditional expectations and conditional covariances.

Proposition A.5.19. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ a standard Borel space, $f : \Omega \to X$ and $g : \Omega \to \mathbb{R}$ random variables, where $\mathsf{E}(|g|) < \infty$, and $\mu : X \to \mathcal{P}(\mathbb{R})$ a regular conditional distribution of g given f. Let $x \in X$. Then, for almost all $\omega \in f^{-1}(\{x\})$,*

$$\mathsf{E}(g \mid f)(\omega) = \int_{\mathbb{R}} \lambda y.y\ d\mu(x).$$

Proof. For almost all $\omega \in f^{-1}(\{x\})$,

$$\begin{aligned}
&\mathsf{E}(g \mid f)(\omega) \\
&= \mathsf{E}(\lambda y.y \circ g \mid f)(\omega) \\
&= \int_{\mathbb{R}} \lambda y.y \; d\mu(f(\omega)) && \text{[Proposition A.5.18, Equation A.5.1]} \\
&= \int_{\mathbb{R}} \lambda y.y \; d\mu(x).
\end{aligned}$$

□

For almost all $\omega \in f^{-1}(\{x\})$, $\mathsf{E}(g \mid f)(\omega)$ is the conditional expectation of g given that f has the value $x \in X$.

Proposition A.5.20. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ a standard Borel space, $f : \Omega \to X$ and $(g_1, g_2) : \Omega \to \mathbb{R}^2$ random variables, where $\mathsf{E}(g_1^2) < \infty$ and $\mathsf{E}(g_2^2) < \infty$, and $\mu : X \to \mathcal{P}(\mathbb{R}^2)$ a regular conditional distribution of (g_1, g_2) given f. Let $x \in X$. Then, for almost all $\omega \in f^{-1}(\{x\})$,*

$$\begin{aligned}
&\mathsf{E}((g_1 - \mathsf{E}(g_1 \mid f)(\omega))\,(g_2 - \mathsf{E}(g_2 \mid f)(\omega)) \mid f)(\omega) = \\
&\quad \int_{\mathbb{R}^2} \lambda(y_1, y_2).(y_1 - \int_{\mathbb{R}} \lambda y.y \; d(\mu(x) \circ \pi_1^{-1}))\,(y_2 - \int_{\mathbb{R}} \lambda y.y \; d(\mu(x) \circ \pi_2^{-1}))\; d\mu(x).
\end{aligned}$$

Proof. For $i = 1, 2$, let $\pi_i \triangleq \lambda(y_1, y_2).y_i : \mathbb{R}^2 \to \mathbb{R}$. Then Proposition A.10.2 below shows that, for $i = 1, 2$, $\lambda x.(\mu(x) \circ \pi_i^{-1}) : X \to \mathcal{P}(\mathbb{R})$ is a regular conditional distribution of g_i given f.

By Hölder's inequality, $\mathsf{E}(|g_1 g_2|) < \infty$. Also $\mathsf{E}(|g_1|) < \infty$ and $\mathsf{E}(|g_2|) < \infty$. Hence, for almost all $\omega \in f^{-1}(\{x\})$, $\mathsf{E}(|(g_1 - \mathsf{E}(g_1 \mid f)(\omega))\,(g_2 - \mathsf{E}(g_2 \mid f)(\omega))|) < \infty$, and so Proposition A.5.18 can be applied below.

For almost all $\omega \in f^{-1}(\{x\})$,

$$\begin{aligned}
&\mathsf{E}((g_1 - \mathsf{E}(g_1 \mid f)(\omega))\,(g_2 - \mathsf{E}(g_2 \mid f)(\omega)) \mid f)(\omega) \\
&= \mathsf{E}(\lambda(y_1, y_2).(y_1 - \mathsf{E}(g_1 \mid f)(\omega))\,(y_2 - \mathsf{E}(g_2 \mid f)(\omega)) \circ (g_1, g_2) \mid f)(\omega) \\
&= \int_{\mathbb{R}^2} \lambda(y_1, y_2).(y_1 - \mathsf{E}(g_1 \mid f)(\omega))\,(y_2 - \mathsf{E}(g_2 \mid f)(\omega))\; d\mu(f(\omega)) \\
&\qquad\qquad \text{[Proposition A.5.18, Equation A.5.1]} \\
&= \int_{\mathbb{R}^2} \lambda(y_1, y_2).(y_1 - \int_{\mathbb{R}} \lambda y.y \; d(\mu(x) \circ \pi_1^{-1}))\,(y_2 - \int_{\mathbb{R}} \lambda y.y \; d(\mu(x) \circ \pi_2^{-1}))\; d\mu(x). \\
&\qquad\qquad \text{[Proposition A.5.19]}
\end{aligned}$$

□

For almost all $\omega \in f^{-1}(\{x\})$, $\mathsf{E}((g_1 - \mathsf{E}(g_1 \mid f)(\omega))\,(g_2 - \mathsf{E}(g_2 \mid f)(\omega)) \mid f)(\omega)$ is the conditional covariance of (g_1, g_2) given that f has the value $x \in X$.

The next result gives a useful connection between the laws of two random variables.

Proposition A.5.21. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ measurable spaces, and $f_1 : \Omega \to X_1$ and $f_2 : \Omega \to X_2$ random variables. Suppose there is a probability kernel $\mu_{1,2} : X_1 \to \mathcal{P}(X_2)$ such that, for all $A_2 \in \mathcal{A}_2$, $\mathsf{P}(f_2^{-1}(A_2) \,|\, f_1) = \lambda\omega.\mu_{1,2}(f_1(\omega))(A_2)$ a.s. Then*

$$(\mathsf{P} \circ f_1^{-1}) \odot \mu_{1,2} = \mathsf{P} \circ f_2^{-1}.$$

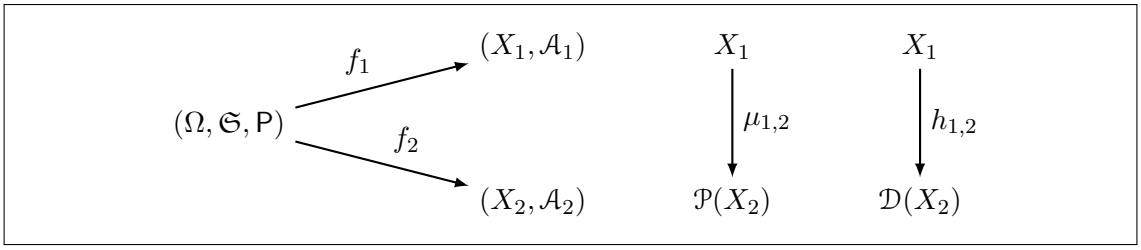

Figure A.9: Setting for Propositions A.5.21 and A.5.24

Proof. For all $A_2 \in \mathcal{A}_2$,

$$\begin{aligned}
&((\mathsf{P} \circ f_1^{-1}) \odot \mu_{1,2})(A_2) \\
&= \int_{X_1} \lambda x_1.\mu_{1,2}(x_1)(A_2)\; d(\mathsf{P} \circ f_1^{-1}) \\
&= \int_{\Omega} \lambda\omega.\mu_{1,2}(f_1(\omega))(A_2)\; d\mathsf{P} \\
&= \int_{\Omega} \mathsf{P}(f_2^{-1}(A_2) \,|\, f_1)\; d\mathsf{P} \\
&= \int_{\Omega} \mathbf{1}_{f_2^{-1}(A_2)}\; d\mathsf{P} \\
&= (\mathsf{P} \circ f_2^{-1})(A_2).
\end{aligned}$$

□

The next result concerns regular conditional distributions on product spaces.

Proposition A.5.22. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_i, \mathcal{A}_i)$ and $(Y_i, \mathcal{B}_i)$ measurable spaces, $f_i : \Omega \to X_i$ and $g_i : \Omega \to Y_i$ random variables, and $\mu_i : X_i \to \mathcal{P}(Y_i)$ a probability kernel, for $i = 1, \ldots, n$. Suppose that*

$$\mathsf{P}(g_i^{-1}(B_i) \,|\, f_i) = \lambda\omega.\mu_i(f_i(\omega))(B_i) \text{ a.s.}$$

and

$$\mathsf{P}(\bigcap_{i=1}^{n} g_i^{-1}(B_i) \,|\, (f_1, \ldots, f_n)) = \prod_{i=1}^{n} \mathsf{P}(g_i^{-1}(B_i) \,|\, f_i) \text{ a.s.},$$

for $i = 1, \ldots, n$ and all $B_i \in \mathcal{B}_i$. Let

$$\mu \triangleq \lambda(x_1, \ldots, x_n).\bigotimes_{i=1}^{n} \mu_i(x_i) : \prod_{i=1}^{n} X_i \to \mathcal{P}(\prod_{i=1}^{n} Y_i).$$

Then, for all $B \in \bigotimes_{i=1}^{n} \mathcal{B}_i$,

$$\mathsf{P}((g_1, \ldots, g_n)^{-1}(B) \,|\, (f_1, \ldots, f_n)) = \lambda\omega.\mu((f_1, \ldots, f_n)(\omega))(B) \text{ a.s.}$$

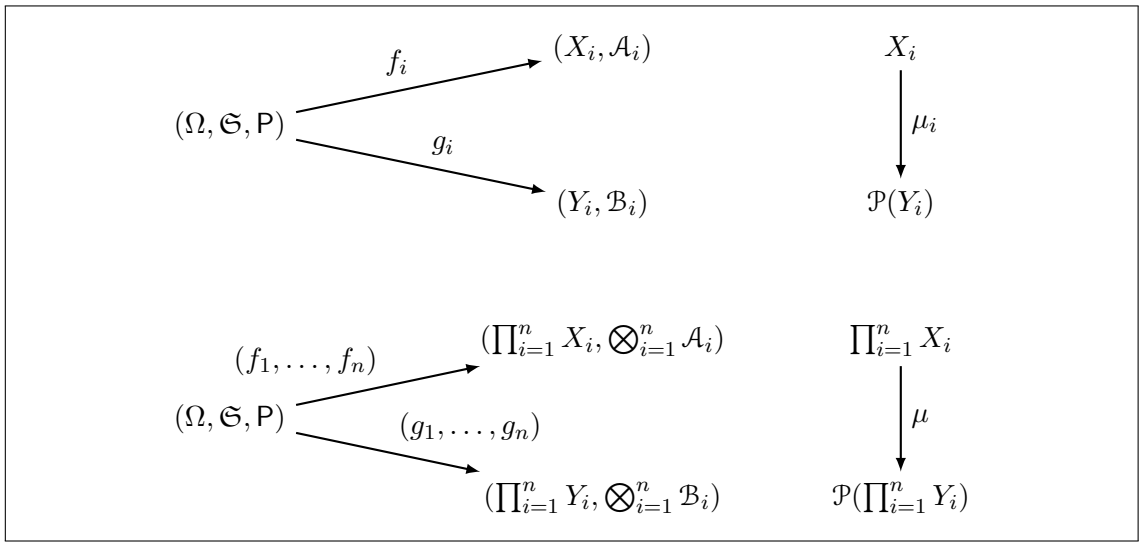

Figure A.10: Setting for Proposition A.5.22

Proof. Proposition A.2.17 shows that μ is a probability kernel.

Let $\mathcal{P} \triangleq \{\prod_{i=1}^n B_i \mid B_i \in \mathcal{B}_i, \text{ for } i = 1, \ldots, n\}$ and

$$\mathcal{L} \triangleq \{B \in \bigotimes_{i=1}^{n} \mathcal{B}_i \mid \mathsf{P}((g_1, \ldots, g_n)^{-1}(B) \,|\, (f_1, \ldots, f_n)) = \lambda\omega.\mu((f_1, \ldots, f_n)(\omega))(B) \text{ a.s.}\}.$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \bigotimes_{i=1}^{n} \mathcal{B}_i$.

Suppose that $\prod_{i=1}^n B_i \in \mathcal{P}$. Then, P-almost surely,

$$\begin{aligned}
&\mathsf{P}((g_1, \ldots, g_n)^{-1}(\prod_{i=1}^{n} B_i) \,|\, (f_1, \ldots, f_n)) \\
&= \prod_{i=1}^{n} \mathsf{P}(g_i^{-1}(B_i) \,|\, f_i) \\
&= \prod_{i=1}^{n} \lambda\omega.\mu_i(f_i(\omega))(B_i) \\
&= \lambda\omega. \bigotimes_{i=1}^{n} \mu_i(f_i(\omega))(\prod_{i=1}^{n} B_i) \\
&= \lambda\omega.\mu((f_1, \ldots, f_n)(\omega))(\prod_{i=1}^{n} B_i).
\end{aligned}$$

Thus $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, $\prod_{i=1}^n X_i \in \mathcal{P}$ and $\mathcal{P} \subseteq \mathcal{L}$, so that $\prod_{i=1}^n X_i \in \mathcal{L}$.

Second, let $A, B \in \mathcal{L}$, where $A \subseteq B$. Then, P-almost surely,

$$\begin{aligned}
&\mathsf{P}((g_1,\ldots,g_n)^{-1}(B \setminus A) \,|\, (f_1,\ldots,f_n)) \\
&= \mathsf{P}((g_1,\ldots,g_n)^{-1}(B) \,|\, (f_1,\ldots,f_n)) - \mathsf{P}((g_1,\ldots,g_n)^{-1}(A) \,|\, (f_1,\ldots,f_n)) \\
&\qquad\qquad \text{[Proposition A.5.6, Part 1]} \\
&= \lambda\omega.\mu((f_1,\ldots,f_n)(\omega))(B) - \lambda\omega.\mu((f_1,\ldots,f_n)(\omega))(A) \\
&= \lambda\omega.\mu((f_1,\ldots,f_n)(\omega))(B \setminus A).
\end{aligned}$$

Hence $B \setminus A \in \mathcal{L}$.

Third, let $(B_k)_{k\in\mathbb{N}}$ be an increasing sequence of sets in $\mathcal{L}$. Then, P-almost surely,

$$\begin{aligned}
&\mathsf{P}((g_1,\ldots,g_n)^{-1}(\bigcup_{k\in\mathbb{N}} B_k) \,|\, (f_1,\ldots,f_n)) \\
&= \lim_{k\to\infty} \mathsf{P}((g_1,\ldots,g_n)^{-1}(B_k) \,|\, (f_1,\ldots,f_n)) \qquad \text{[Proposition A.5.6, Part 4]} \\
&= \lim_{k\to\infty} \lambda\omega.\mu((f_1,\ldots,f_n)(\omega))(B_k) \\
&= \lambda\omega.\mu((f_1,\ldots,f_n)(\omega))(\bigcup_{k\in\mathbb{N}} B_k). \qquad \text{[Proposition A.2.2]}
\end{aligned}$$

Hence $\bigcup_{k\in\mathbb{N}} B_k \in \mathcal{L}$.

It now follows from the monotone-class theorem (Proposition A.1.2) that $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. Hence the result. □

Proposition A.5.22 states that, under a conditional independence assumption, if each μ_i is a regular conditional distribution of g_i given f_i, then μ is a regular conditional distribution of $(g_1,\ldots,g_n)$ given $(f_1,\ldots,f_n)$.

Here is a companion to Proposition A.5.22 for a different, but similar, setting.

Proposition A.5.23. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_i, \mathcal{A}_i)$ and $(Y, \mathcal{B})$ measurable spaces, $f_i : \Omega \to X_i$ and $g : \Omega \to Y$ random variables, and $\mu_i : X_i \to \mathcal{P}(Y)$ a probability kernel, for $i = 1,\ldots,n$. Suppose that*

$$\mathsf{P}(g^{-1}(B) \,|\, f_i) = \lambda\omega.\mu_i(f_i(\omega))(B) \text{ a.s.}$$

and

$$\mathsf{P}(g^{-1}(B) \,|\, (f_1,\ldots,f_n)) = \prod_{i=1}^n \mathsf{P}(g^{-1}(B) \,|\, f_i) \text{ a.s.},$$

for $i = 1,\ldots,n$ and all $B \in \mathcal{B}$. Let

$$\mu \triangleq \lambda(x_1,\ldots,x_n).\lambda B.\prod_{i=1}^n \mu_i(x_i)(B) : \prod_{i=1}^n X_i \to \mathcal{P}(Y).$$

Then, for all $B \in \mathcal{B}$,

$$\mathsf{P}(g^{-1}(B) \,|\, (f_1,\ldots,f_n)) = \lambda\omega.\mu((f_1,\ldots,f_n)(\omega))(B) \text{ a.s.}$$

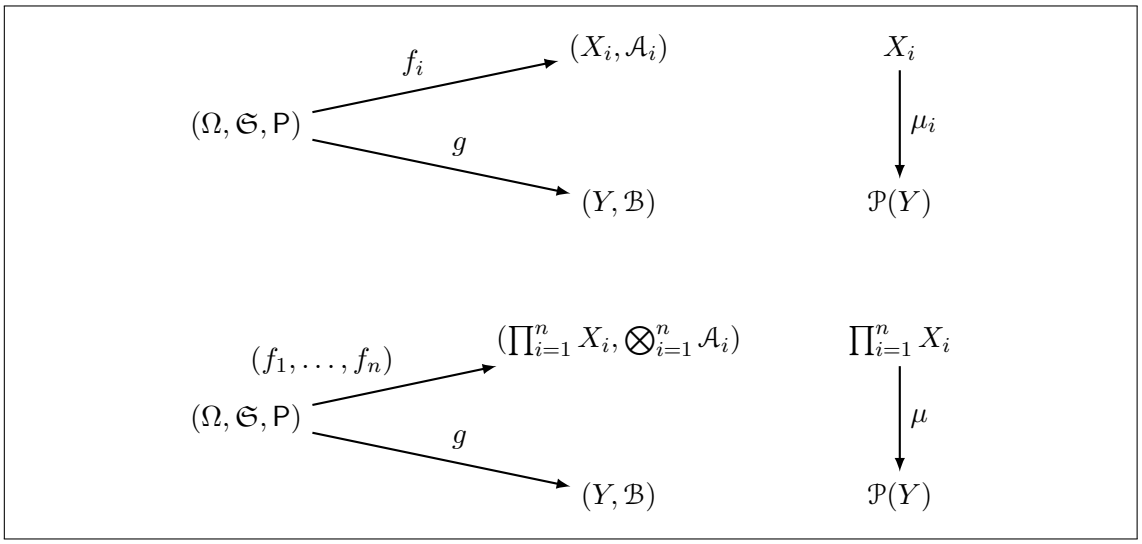

Figure A.11: Setting for Proposition A.5.23

Proof. Using Proposition A.2.4, it can be shown that μ is a probability kernel.

Suppose that $B \in \mathcal{B}$. Then, P-almost surely,

$$\begin{aligned}
&\mathsf{P}(g^{-1}(B) \,|\, (f_1, \ldots, f_n)) \\
&= \prod_{i=1}^n \mathsf{P}(g^{-1}(B) \,|\, f_i) \\
&= \prod_{i=1}^n \lambda\omega.\mu_i(f_i(\omega))(B) \\
&= \lambda\omega. \prod_{i=1}^n \mu_i(f_i(\omega))(B) \\
&= \lambda\omega.\mu((f_1, \ldots, f_n)(\omega))(B).
\end{aligned}$$

□

Proposition A.5.23 states that, under a conditional independence assumption, if each μ_i is a regular conditional distribution of g given f_i, then μ is a regular conditional distribution of g given $(f_1, \ldots, f_n)$.

Here is the analogue of Proposition A.5.21 for conditional densities.

Proposition A.5.24. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_1, \mathcal{A}_1, \nu_1)$ and $(X_2, \mathcal{A}_2, \nu_2)$ σ-finite measure spaces, and $f_1 : \Omega \to X_1$ and $f_2 : \Omega \to X_2$ random variables. Suppose there exists $h_1 \in \mathcal{D}(X_1)$ such that $\mathsf{P} \circ f_1^{-1} = h_1 \cdot \nu_1$, $h_2 \in \mathcal{D}(X_2)$ such that $\mathsf{P} \circ f_2^{-1} = h_2 \cdot \nu_2$, and a conditional density $h_{1,2} : X_1 \to \mathcal{D}(X_2)$ such that, for all $A_1 \in \mathcal{A}_2$, $\mathsf{P}(f_2^{-1}(A_2) \,|\, f_1) = \lambda\omega.(h_{1,2} \cdot \nu_2)(f_1(\omega))(A_2)$ a.s. Then*

$$h_1 \odot h_{1,2} = h_2 \;\; \nu_2\text{-a.e.}$$

Proof. For all $A_2 \in \mathcal{A}_2$,

$$\begin{aligned}
&\int_{X_2} \mathbf{1}_{A_2}\, h_1 \odot h_{1_2}\, d\nu_2 \\
&= \int_{X_2} \mathbf{1}_{A_2} \left(\lambda x_2. \int_{X_1} \lambda x_1. h_{1,2}(x_1)(x_2)\, h_1(x_1)\, d\nu_1 \right)\, d\nu_2 \\
&= \int_{X_2} \left(\lambda x_2. \int_{X_1} \lambda x_1. \mathbf{1}_{A_2}(x_2)\, h_{1,2}(x_1)(x_2)\, h_1(x_1)\, d\nu_1 \right)\, d\nu_2 \\
&= \int_{X_1} \left(\lambda x_1. \int_{X_2} \lambda x_2. \mathbf{1}_{A_2}(x_2)\, h_{1,2}(x_1)(x_2)\, h_1(x_1)\, d\nu_2 \right)\, d\nu_1 \\
&= \int_{X_1} \lambda x_1. \left(\int_{X_2} \mathbf{1}_{A_2}\, h_{1,2}(x_1)\, d\nu_2 \right) h_1(x_1)\, d\nu_1 \\
&= \int_{X_1} \lambda x_1. \left(\int_{X_2} \mathbf{1}_{A_2}\, h_{1,2}(x_1)\, d\nu_2 \right)\, d(h_1 \cdot \nu_1) \\
&= \int_{X_1} \lambda x_1. \left(\int_{X_2} \mathbf{1}_{A_2}\, h_{1,2}(x_1)\, d\nu_2 \right)\, d(\mathsf{P} \circ f_1^{-1}) \\
&= \int_{X_1} \lambda x_1. (h_{1,2} \cdot \nu_2)(x_1)(A_2)\, d(\mathsf{P} \circ f_1^{-1}) \\
&= \int_{\Omega} \lambda \omega. (h_{1,2} \cdot \nu_2)(f_1(\omega))(A_2)\, d\mathsf{P} \\
&= \int_{\Omega} \mathsf{P}(f_2^{-1}(A_2) \,|\, f_1)\, d\mathsf{P} \\
&= \int_{\Omega} \mathbf{1}_{f_2^{-1}(A_2)}\, d\mathsf{P} \\
&= (\mathsf{P} \circ f_2^{-1})(A_2) \\
&= (h_2 \cdot \nu_2)(A_2) \\
&= \int_{X_2} \mathbf{1}_{A_2}\, h_2\, d\nu_2.
\end{aligned}$$

It follows from Proposition A.2.11 that $h_1 \odot h_{1,2} = h_2$ ν_2-a.e. □

Here is the analogue of Definition A.5.12 for conditional densities.

Definition A.5.13. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(Y, \mathcal{B})$, a measurable space, $(Z, \mathcal{C}, \nu)$ a σ-finite measure space, and $f : \Omega \to Y$ and $g : \Omega \to Z$ random variables. A conditional density $h : Y \to \mathcal{D}(Z)$ is called a *regular conditional density* (*with respect to* ν) of g given f if, for all $C \in \mathcal{C}$,

$$\mathsf{P}(g^{-1}(C) \,|\, f) = \lambda \omega. (h \cdot \nu)(f(\omega))(C) \text{ a.s.}$$

In other words, h is a regular conditional density (with respect to ν) iff $h \cdot \nu$ is a regular conditional distribution.

A.6 Conditional Independence

Throughout, conditional independence assumptions play an important role.

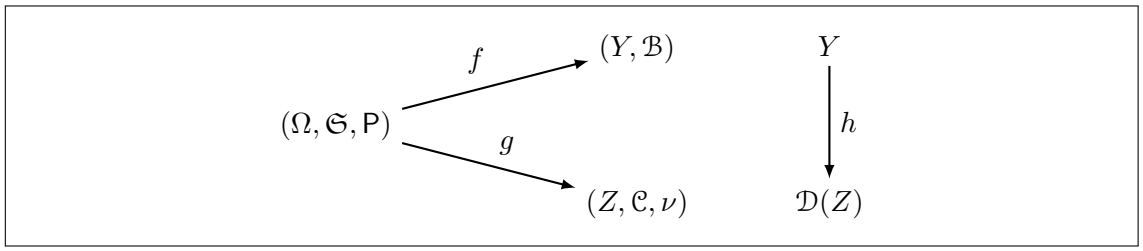

Figure A.12: Setting for Definition A.5.13

Definition A.6.1. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ sub-σ-algebras of $\mathfrak{S}$. Then $\mathcal{F}$ and $\mathcal{H}$ are *conditionally independent given* $\mathcal{G}$, denoted by

$$\mathcal{F} \underset{\mathcal{G}}{\perp\!\!\!\perp} \mathcal{H},$$

if

$$\mathsf{P}(F \cap H \,|\, \mathcal{G}) = \mathsf{P}(F \,|\, \mathcal{G})\ \mathsf{P}(H \,|\, \mathcal{G}) \text{ a.s.},$$

for all $F \in \mathcal{F}$ and $H \in \mathcal{H}$.

Note that conditional independence reduces to independence if $\mathcal{G}$ is the trivial σ-algebra $\{\emptyset, \Omega\}$.

Notation. If $\mathcal{F}$ and $\mathcal{G}$ are σ-algebras, then $\sigma(\mathcal{F}, \mathcal{G})$ denotes the smallest σ-algebra containing $\mathcal{F}$ and $\mathcal{G}$.

If $\mathcal{F}_1, \mathcal{F}_2, \ldots$ are σ-algebras, then $\sigma(\mathcal{F}_1, \mathcal{F}_2, \ldots)$ denotes the smallest σ-algebra containing $\mathcal{F}_i$, for $i = 1, 2, \ldots$.

If $(\mathcal{F}_i)_{i \in I}$ is an indexed family of σ-algebras, then $\mathcal{F}_I$ denotes the smallest σ-algebra containing $\mathcal{F}_i$, for $i \in I$. (In other words, $\mathcal{F}_I$ is shorthand for $\sigma((\mathcal{F}_i)_{i \in I})$.)

Note that, if I is empty, then $\mathcal{F}_I$ is the trivial σ-algebra $\{\emptyset, \Omega\}$.

A useful characterisation of conditional independence is given by the next result.

Proposition A.6.1. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ sub-σ-algebras of $\mathfrak{S}$. Then $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$ iff*

$$\mathsf{P}(H \,|\, \sigma(\mathcal{F}, \mathcal{G})) = \mathsf{P}(H \,|\, \mathcal{G}) \text{ a.s.},$$

for all $H \in \mathcal{H}$.

Proof. See [83, Proposition 6.6] □

Proposition A.6.2. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ sub-σ-algebras of $\mathfrak{S}$. Then $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$ iff $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \sigma(\mathcal{G}, \mathcal{H})$.*

Proof. See [83, Corollary 6.7]. □

Proposition A.6.3. *(Chain Rule) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $\mathcal{G}$, $\mathcal{H}$, and $\mathcal{F}_1, \mathcal{F}_2, \ldots$ sub-σ-algebras of $\mathfrak{S}$. Then the following are equivalent.*

1. $\mathcal{H} \underset{\mathcal{G}}{\perp\!\!\!\perp} \sigma(\mathcal{F}_1, \mathcal{F}_2, \ldots)$.
2. $\mathcal{H} \underset{\sigma(\mathcal{G}, \mathcal{F}_1, \ldots, \mathcal{F}_n)}{\perp\!\!\!\perp} \mathcal{F}_{n+1}$ *for* $n = 0, 1, \ldots$.

In particular,

$$\mathcal{H} \underset{\mathcal{G}}{\perp\!\!\!\perp} \sigma(\mathcal{F}_1, \mathcal{F}_2) \quad \textit{iff} \quad \mathcal{H} \underset{\mathcal{G}}{\perp\!\!\!\perp} \mathcal{F}_1 \ \textit{ and } \ \mathcal{H} \underset{\sigma(\mathcal{G}, \mathcal{F}_1)}{\perp\!\!\!\perp} \mathcal{F}_2.$$

Proof. See [83, Proposition 6.8]. □

For finding implicit conditional independencies, the concept of a dependency graph will be central.

Definition A.6.2. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space. A *dependency graph* (*with respect to* $(\Omega, \mathfrak{S}, \mathsf{P})$) is a directed acyclic graph whose vertices are labelled by sub-σ-algebras of $\mathfrak{S}$.

Positive integers are used to denote vertices in a dependency graph. Intuitively, an edge that is directed from vertex i to vertex j indicates that i directly influences j. Dependency graphs are finite; for the applications considered in this book, dependency graphs can increase in size with each time step, but there is never any need to consider infinite dependency graphs. For vertex i, the σ-algebra that labels the vertex is is denoted by $\mathcal{F}_i$. Distinct vertices may be labelled by the same σ-algebra. If a vertex is labelled by a σ-algebra $\mathcal{F}$ and $\mathcal{F} = \sigma(f)$, for some random variable f, then the vertex may be labelled by f instead.

Definition A.6.3. A *topological order* on a dependency graph is a total order $\prec$ on the set of vertices such that, for all vertices i and j, if i is parent of j, then $i \prec j$.

For an example of a topological order, see Figure A.13.

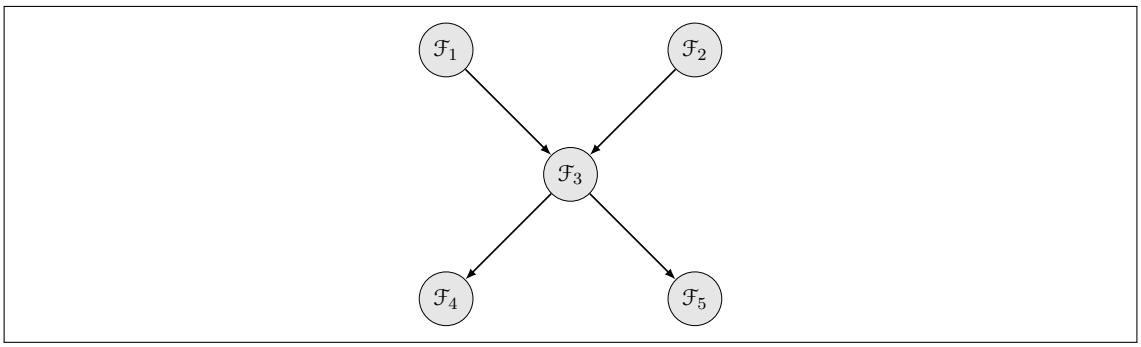

Figure A.13: A dependency graph such that $1, 2, 3, 4, 5$ is a topological order of the vertices

By renaming vertices if necessary, the assumption can be made that, without loss of generality, the vertices are denoted by $1, \ldots, n$, where n is the number of vertices, and the usual order $<$ on integers is a topological order of the vertices. This is done in the following discussion.

Notation. For a vertex i, let $par(i)$ denote the set of parent vertices of i in the graph.

Note that, since $<$ is assumed to be a topological order of the vertices, it follows that $par(i) \subseteq \{1, \ldots, i-1\}$, for $i = 1, \ldots, n$.

The next concept is used to find implicit conditional independencies.

Definition A.6.4. Let A, B, and C be pairwise disjoint subsets of vertices in a dependency graph. Then a path from a vertex in A to a vertex in B is *blocked* by C if at least on of the following holds:

1. There is a vertex on the path where the edges meet either head-to-tail or tail-to-tail and the vertex is in C, or

2. There is a vertex on the path where the edges meet head-to-head and neither the vertex nor any of its descendants are in C.

A and B are *d-separated* by C if all paths from a vertex in A to a vertex in B are blocked by C.

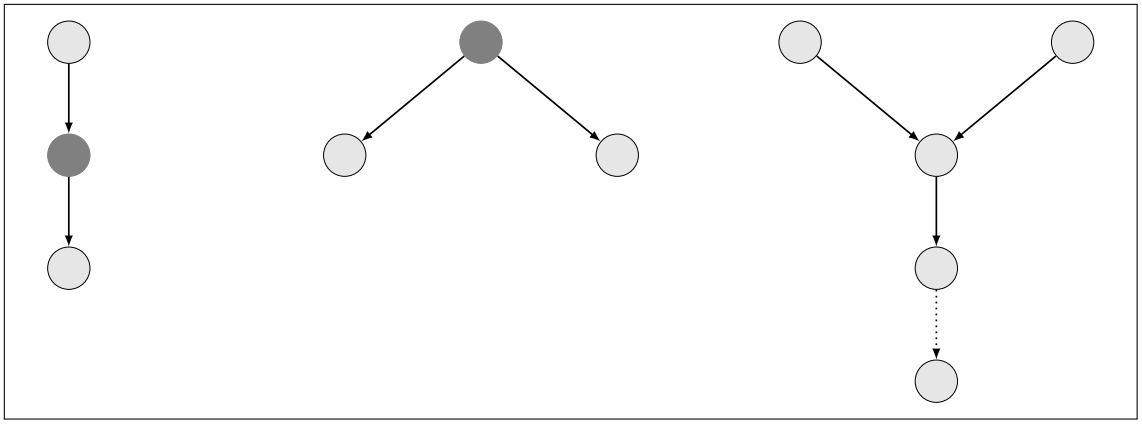

Figure A.14: In the leftmost and middle graphs, the path is blocked at the observed dark-coloured vertex. In the rightmost graph, the path is blocked at the unobserved vertex where the edges meet head-to-head

Note that the definition of d-separated is symmetric in A and B. Vertices in C are traditionally called *observed* vertices. Clearly, A and B are d-separated by C iff all paths that do not contains loops from a vertex in A to a vertex in B are blocked by C.

As will be seen later in the context of probability measures on product spaces, a dependency graph may satisfy certain conditional independencies introduced in the following definition. Recall that, unless stated otherwise, it is assumed throughout that $1, \ldots, n$ is a topological order of the set of n vertices of a dependency graph.

Definition A.6.5. A dependency graph is *Markov* if

$$\mathcal{F}_i \underset{\mathcal{F}_{par(i)}}{\perp\!\!\!\perp} \mathcal{F}_{\{1,\ldots,i-1\}},$$

for $i = 1, \ldots, n$.

By Proposition A.6.1, a dependency graph is Markov iff, for $i = 1, \ldots, n$,

$$\mathsf{P}(F \mid \mathcal{F}_{\{1,\ldots,i-1\}}) = \mathsf{P}(F \mid \mathcal{F}_{par(i)}) \text{ a.s.},$$

for all $F \in \mathcal{F}_i$.

Note that the Markov property in Definition A.6.5 appears to depend upon the topological order chosen since the set $\{1, \ldots, i-1\}$ depends upon this order. This will clarified in Proposition A.6.6.

In the case where each vertex has an edge to it from each of its predecessors, $par_i = \{1, \ldots, i-1\}$, for all $i, \ldots, n$, and hence the dependency graph is Markov no matter what σ-algebras label its vertices. The interesting case is where par_i is a strict subset of $\{1, \ldots, i-1\}$, for at least some values of i. From a practical point of view, it is best that each $|par_i|$ be as small as possible.

Example A.6.1. Consider the dependency graph shown in Figure A.15. For this graph, the conditions to be Markov are as follows:

$$\mathcal{F}_1 \underset{\{\emptyset,\Omega\}}{\perp\!\!\!\perp} \{\emptyset, \Omega\}, \qquad \mathcal{F}_2 \underset{\mathcal{F}_1}{\perp\!\!\!\perp} \mathcal{F}_1, \qquad \mathcal{F}_3 \underset{\mathcal{F}_2}{\perp\!\!\!\perp} \mathcal{F}_{\{1,2\}}, \qquad \mathcal{F}_4 \underset{\mathcal{F}_2}{\perp\!\!\!\perp} \mathcal{F}_{\{1,2,3\}}.$$

Note that $\mathcal{F}_1 \perp\!\!\!\perp_{\{\emptyset,\Omega\}} \{\emptyset, \Omega\}$ and $\mathcal{F}_2 \perp\!\!\!\perp_{\mathcal{F}_1} \mathcal{F}_1$ are true, no matter what the choice of $\mathcal{F}_1$ and $\mathcal{F}_2$.

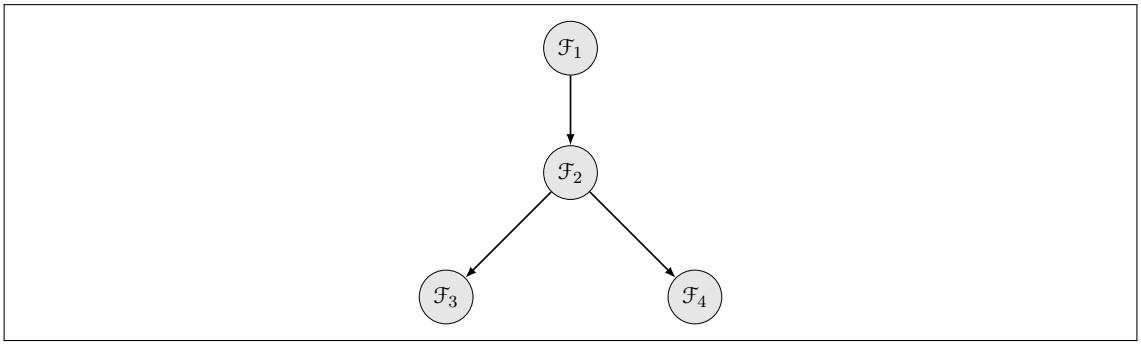

Figure A.15: A dependency graph showing conditional independencies

The next result shows that the Markov condition implies many more implicit conditional independencies.

Proposition A.6.4. *Let G be a Markov dependency graph and A, B, and C pairwise disjoint subsets of vertices of G such that A and B are d-separated by C. Then*

$$\mathcal{F}_A \underset{\mathcal{F}_C}{\perp\!\!\!\perp} \mathcal{F}_B.$$

Proof. The proof is by induction on the number of vertices n in G. If $n = 1$, then at least one of A and B is empty, and so at least one of $\mathcal{F}_A$ and $\mathcal{F}_B$ is the trivial σ-algebra. Thus $\mathcal{F}_A \perp\!\!\!\perp_{\mathcal{F}_C} \mathcal{F}_B$.

Now suppose the result holds for graphs of n vertices and consider a graph G containing $n+1$ vertices that satisfies $\mathcal{F}_i \perp\!\!\!\perp_{\mathcal{F}_{par(i)}} \mathcal{F}_{\{1,\ldots,i-1\}}$, for $i = 1, \ldots, n+1$. The vertex labelled

$n+1$ in G is last in the topological order and hence has no children. Consider the graph G' obtained from G by removing vertex $n+1$ (and any edges leading to it). Note that G' satisfies $\mathcal{F}_i \perp\!\!\!\perp_{\mathcal{F}_{par(i)}} \mathcal{F}_{\{1,\ldots,i-1\}}$, for $i = 1,\ldots,n$. Let A, B and C be pairwise disjoint subsets of $\{1,\ldots,n+1\}$ such that A and B are d-separated by C (in G). There are three cases to consider.

(a) Suppose that $n+1 \notin A \cup B \cup C$. Then A and B are d-separated by C (in G'), since the only contribution $n+1$ can make towards blocking paths is as a descendant of an unobserved vertex where the edges meets head-to-head and deleting $n+1$ still leaves the path blocked at the unobserved vertex. Hence, by the induction hypothesis, $\mathcal{F}_A \perp\!\!\!\perp_{\mathcal{F}_C} \mathcal{F}_B$.

(b) Suppose that $n+1 \in A$. (The case when $n+1 \in B$ is handled in a similar way.) Let $A' \triangleq A \setminus \{n+1\}$. Then A' and B and d-separated by C (in G'). Since there can be no edges from A to B, it follows that no parent of $n+1$ is in B. Let P be the parents of $n+1$ that are not in C. It is shown that P and B are d-separated by C (in G').

Consider any path (with no loops) from P to B (in G'). Such a path can be extended to a path from $n+1$ to B by adding the edge from the appropriate parent in P to $n+1$. Since A and B are d-separated by C, this extended path must be blocked in G and, since it cannot be blocked at $n+1$ or its (unobserved) parent, it must be blocked in G'. Hence the path from P to B (in G') is blocked. It follows that P and B are d-separated by C (in G').

Thus $A' \cup P$ and B are separated by C (in G'). By the induction hypothesis, $\mathcal{F}_{A'\cup P} \perp\!\!\!\perp_{\mathcal{F}_C} \mathcal{F}_B$.

Next it is shown that $\mathcal{F}_{n+1} \perp\!\!\!\perp_{\mathcal{F}_{A'\cup C\cup P}} \mathcal{F}_B$. For this, note that $par(n+1) \subseteq C \cup P$. Now $\mathcal{F}_{n+1} \perp\!\!\!\perp_{\mathcal{F}_{par(n+1)}} \mathcal{F}_{B\cup((A'\cup C\cup P)\setminus par(n+1))}$, since $\mathcal{F}_{n+1} \perp\!\!\!\perp_{\mathcal{F}_{par(n+1)}} \mathcal{F}_{\{1,\ldots,n\}}$. By Proposition A.6.3, it follows that $\mathcal{F}_{n+1} \perp\!\!\!\perp_{\mathcal{F}_{A'\cup C\cup P}} \mathcal{F}_B$.

By Proposition A.6.3, $\mathcal{F}_{A'\cup P} \perp\!\!\!\perp_{\mathcal{F}_C} \mathcal{F}_B$ and $\mathcal{F}_{n+1} \perp\!\!\!\perp_{\mathcal{F}_{A'\cup C\cup P}} \mathcal{F}_B$ imply $\mathcal{F}_{A\cup P} \perp\!\!\!\perp_{\mathcal{F}_C} \mathcal{F}_B$. Hence $\mathcal{F}_A \perp\!\!\!\perp_{\mathcal{F}_C} \mathcal{F}_B$.

(c) Suppose that $n+1 \in C$. In this case, no path can be blocked at $n+1$; hence A and B must be d-separated by $C' \triangleq C \setminus \{n+1\}$. Also $\{n+1\}$ must be d-separated from A or B or both by C', otherwise there would be a path from A to B via $n+1$ that is not blocked in C'. Suppose that this holds for B (the case for A is similar), so that $A \cup \{n+1\}$ and B are d-separated by C'. By Part (b) above, it follows that $\mathcal{F}_{A\cup\{n+1\}} \perp\!\!\!\perp_{\mathcal{F}_{C'}} \mathcal{F}_B$. Then, by Proposition A.6.3, $\mathcal{F}_A \perp\!\!\!\perp_{\mathcal{F}_{C'\cup\{n+1\}}} \mathcal{F}_B$, so that $\mathcal{F}_A \perp\!\!\!\perp_{\mathcal{F}_C} \mathcal{F}_B$. □

Notation. For $i = 1,\ldots,n$, let $nondesc(i)$ denote the set of vertices that are not descendants of vertex i.

Proposition A.6.5. *For a Markov dependency graph,*

$$\mathcal{F}_i \perp\!\!\!\perp_{\mathcal{F}_{par(i)}} \mathcal{F}_{nondesc(i)},$$

for $i = 1,\ldots,n$.

Proof. For $i = 1,\ldots,n$, consider any path from i to a vertex in $nondesc(i) \setminus par(i)$. Such a path must either contain a parent of i where the edges meet head-to-tail or tail-to-tail or else contain a descendant of i where the edges meet head-to-head. In the first case, the path is blocked at the parent of i. In the second case, the path is blocked at the descendant of i. Hence $\{i\}$ and $nondesc(i) \setminus par(i)$ are d-separated by $par(i)$, for $i = 1,\ldots,n$. Thus,

according to Proposition A.6.4,

$$\mathcal{F}_i \underset{\mathcal{F}_{par(i)}}{\perp\!\!\!\perp} \mathcal{F}_{nondesc(i)\setminus par(i)},$$

for $i = 1, \ldots, n$. Hence, by Proposition A.6.2,

$$\mathcal{F}_i \underset{\mathcal{F}_{par(i)}}{\perp\!\!\!\perp} \mathcal{F}_{nondesc(i)},$$

for $i = 1, \ldots, n$. □

Note that the condition in Proposition A.6.5 does not depend on the topological order on the vertices of the dependency graph, since $par(i)$ and $nondesc(i)$ depend only on structural properties of the dependency graph, not the order chosen.

Example A.6.2. Consider the dependency graph in Figure A.16. Then, for example, the Markov property gives the condition

$$\mathcal{F}_4 \underset{\mathcal{F}_{\{1,2\}}}{\perp\!\!\!\perp} \mathcal{F}_{\{1,2,3\}},$$

whereas Proposition A.6.5 gives the condition

$$\mathcal{F}_4 \underset{\mathcal{F}_{\{1,2\}}}{\perp\!\!\!\perp} \mathcal{F}_{\{1,2,3,5\}}.$$

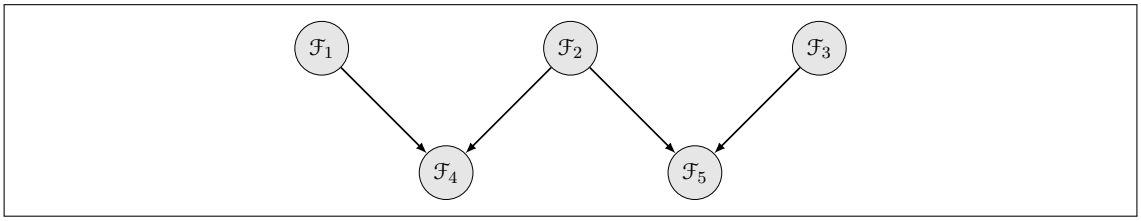

Figure A.16: Dependency graph for Example A.6.2

The Markov property of Definition A.6.5 appears to depend on the choice of topological ordering. However, an important result is that, if a dependency graph is Markov with respect to one topological order, then it is Markov with respect to any other topological order.

Proposition A.6.6. *Suppose that a dependency graph is Markov with respect to some topological order of its vertices. Then it is Markov with respect to any other topological order.*

Proof. Let $<$ and $<'$ be any two topological orders of the vertices in the graph. Suppose the dependency graph is Markov for the topological order $<$. Without loss of generality, it can be assumed that the topological order for $<'$ is the usual order on integers, so that $1, \ldots, n$ is the order of the vertices under $<'$.

By Proposition A.6.5, using the topological order $<$, it follows that

$$\mathcal{F}_i \underset{\mathcal{F}_{par(i)}}{\perp\!\!\!\perp} \mathcal{F}_{nondesc(i)},$$

for $i = 1, \ldots, n$. Whatever the choice of $<'$,

$$\mathcal{F}_{\{1,\ldots,i-1\}} \subseteq \mathcal{F}_{nondesc(i)},$$

for $i = 1, \ldots, n$. Hence

$$\mathcal{F}_i \underset{\mathcal{F}_{par(i)}}{\perp\!\!\!\perp} \mathcal{F}_{\{1,\ldots,i-1\}},$$

for $i = 1, \ldots, n$. Hence the dependency graph is Markov for the topological order $<'$. □

Since the d-separation condition of Proposition A.6.4 is independent of the choice of topological order, Proposition A.6.6 implies that one can use any topological order to verify the Markov property for the purpose of discovering conditional independencies via d-separation for a particular dependency graph.

Definition A.6.6. For $i = 1, \ldots, n$, the *Markov blanket* of a vertex i is the set its parents, children, and children's parents.

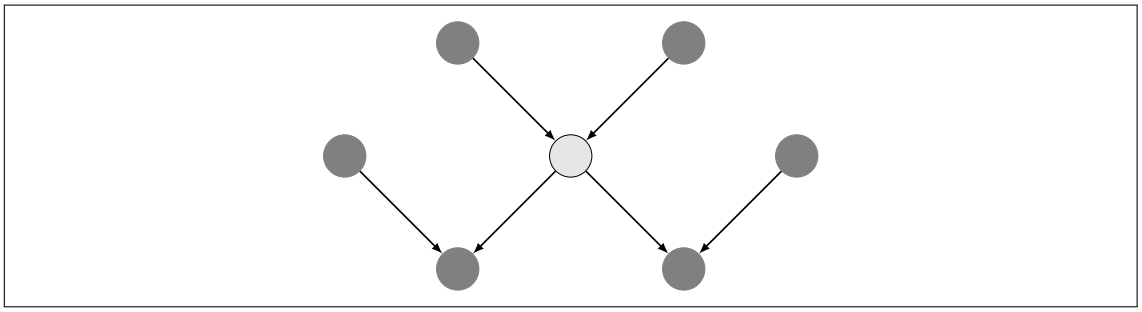

Figure A.17: The Markov blanket for the central vertex consists of the dark-coloured vertices

Notation. For $i = 1, \ldots, n$, let $blanket(i)$ denote the Markov blanket of vertex i.

Proposition A.6.7. *For a Markov dependency graph,*

$$\mathcal{F}_i \underset{\mathcal{F}_{blanket(i)}}{\perp\!\!\!\perp} \mathcal{F}_{\{1,\ldots,i-1,i+1,\ldots,n\}},$$

for $i = 1, \ldots, n$.

Proof. Let $remain(i) \triangleq \{1, \ldots, i-1, i+1, \ldots, n\} \setminus blanket(i)$. For $i = 1, \ldots, n$, consider any path from i to a vertex in $remain(i)$. There are three cases to consider. In the first case, the path contains a parent of i where the edges meet head-to-tail or tail-to-tail, in which case the path is blocked at the parent. In the second case, the path contains a child of i where the edges meet head-to-tail, in which case the path is blocked at the child

vertex. In the third case, the path contains a parent of a child of i where the edges meet head-to-tail or tail-to-tail, in which case the path is blocked at the parent of the child. Hence $\{i\}$ and $remain(i)$ are d-separated by $blanket(i)$, for $i = 1, \ldots, n$. Thus, according to Proposition A.6.4,

$$\mathcal{F}_i \underset{\mathcal{F}_{blanket(i)}}{\perp\!\!\!\perp} \mathcal{F}_{remain(i)},$$

for $i = 1, \ldots, n$. Hence, by Proposition A.6.2,

$$\mathcal{F}_i \underset{\mathcal{F}_{blanket(i)}}{\perp\!\!\!\perp} \mathcal{F}_{\{1,\ldots,i-1,i+1,\ldots,n\}},$$

for $i = 1, \ldots, n$. □

Later on, it is shown that (under weak conditions) a probability measure on a product space can be deconstructed into a product of probability kernels in a way consistent with the dependency graph giving the conditional independencies. (See Proposition A.7.23.)

Here is a useful result about regular conditional distributions under certain conditional independence assumptions.

Proposition A.6.8. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ measurable spaces, $f : \Omega \to X$, $g : \Omega \to Y$, and $h : \Omega \to Z$ random variables, and $\mu : X \to \mathcal{P}(Z)$ a probability kernel that is a regular conditional distribution of h given f. Suppose that $\sigma(g) \perp\!\!\!\perp_{\sigma(f)} \sigma(h)$. Then the probability kernel $\lambda(x, y).\mu(x) : X \times Y \to \mathcal{P}(Z)$ is a regular conditional distribution of h given (f, g).*

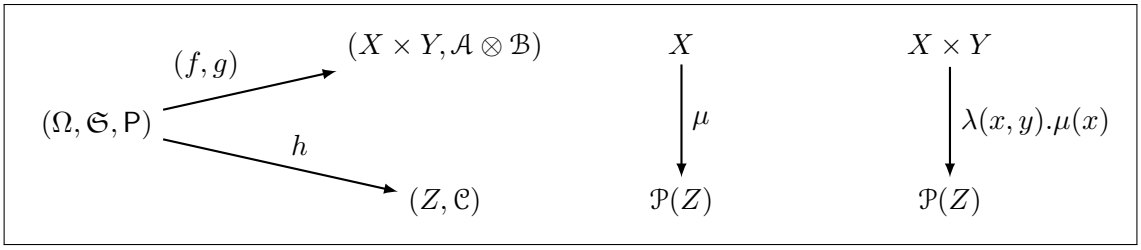

Figure A.18: Setting for Proposition A.6.8

Proof. Since $\mu : X \to \mathcal{P}(Z)$ is a regular conditional distribution of h given f, it follows that $\mathsf{P}(h^{-1}(C) \mid f) = \lambda\omega.\mu(f(\omega))(C)$ a.s., for all $C \in \mathcal{C}$. Also, since $\sigma(g) \perp\!\!\!\perp_{\sigma(f)} \sigma(h)$, $\mathsf{P}(h^{-1}(C) \mid f) = \mathsf{P}(h^{-1}(C) \mid (f, g))$ a.s., for all $C \in \mathcal{C}$.

Now $\lambda(x, y).\mu(x)$ is a probability kernel being the composition of the projection from $X \times Y$ onto X and μ, and $\lambda\omega.(\lambda(x, y).\mu(x))((f, g)(\omega)) = \lambda\omega.\mu(f(\omega))$. Thus

$$\mathsf{P}(h^{-1}(C) \mid (f, g)) = \lambda\omega.(\lambda(x, y).\mu(x))((f, g)(\omega))(C) \text{ a.s.},$$

for all $C \in \mathcal{C}$. Hence $\lambda(x, y).\mu(x)$ is a regular conditional distribution of h given (f, g). □

A.7 Finite Products of Probability Kernels

Much use will be made of the products of finitely many probability kernels.

Definition A.7.1. Let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ be measurable spaces, and $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels. Then the *product* $\mu_1 \otimes \mu_2 : X_0 \to \mathcal{P}(X_1 \times X_2)$ of μ_1 and μ_2 is defined by

$$(\mu_1 \otimes \mu_2)(x_0) = \lambda A. \int_{X_1} \left(\lambda x_1. \int_{X_2} \lambda x_2. \mathbf{1}_A(x_1, x_2)\, d\mu_2(x_0, x_1) \right) d\mu_1(x_0),$$

for all $x_0 \in X_0$. Each of μ_1 and μ_2 is called a *factor* of the product.

Of course, in Definition A.7.1, A ranges over all elements (that is, measurable sets) in $\mathcal{A}_1 \otimes \mathcal{A}_2$.

The product is indeed a probability kernel.

Proposition A.7.1. *Let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ be measurable spaces, and $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels. Then $\mu_1 \otimes \mu_2 : X_0 \to \mathcal{P}(X_1 \times X_2)$ is a probability kernel. Furthermore, $\mu_1 \otimes \mu_2$ is uniquely defined by the property:*

$$(\mu_1 \otimes \mu_2)(x_0)(A_1 \times A_2) = \int_{X_1} \mathbf{1}_{A_1}\, \lambda x_1. \mu_2(x_0, x_1)(A_2)\, d\mu_1(x_0),$$

for all $x_0 \in X_0$, $A \in \mathcal{A}_1$, and $A_2 \in \mathcal{A}_2$.

Proof. See [87, Theorem 14.22]. □

Note. Products and fusions of probability kernels are closely related:

$$(\mu_1 \odot \mu_2)(x_0)(A_2) = (\mu_1 \otimes \mu_2)(x_0)(X_1 \times A_2),$$

for all $x_0 \in X_0$ and $A_2 \in \mathcal{A}_2$. From this, we have

$$\begin{aligned} &(\mu_1 \odot \mu_2)(x_0)(A_2) \\ &= (\mu_1 \otimes \mu_2)(x_0)(X_1 \times A_2) \\ &= (\mu_1 \otimes \mu_2)(x_0)(\pi_2^{-1}(A_2)), \end{aligned}$$

for all $x_0 \in X_0$ and $A_2 \in \mathcal{A}_2$, where $\pi_2 : X_1 \times X_2 \to X_2$ is the canonical projection. Hence

$$\mu_1 \odot \mu_2 = \lambda x_0. ((\mu_1 \otimes \mu_2)(x_0) \circ \pi_2^{-1}).$$

Thus $\mu_1 \odot \mu_2$ is the marginal probability kernel for $\mu_1 \otimes \mu_2$ with respect to X_2. (See Definition A.2.9.) In addition,

$$\begin{aligned} &(\mu_1 \otimes \mu_2)(x_0)(\pi_1^{-1}(A_1)) \\ &= (\mu_1 \otimes \mu_2)(x_0)(A_1 \times X_2) \\ &= \int_{X_1} \mathbf{1}_{A_1} \lambda x_1. \mu_2(x_0, x_1)(X_2)\, d\mu_1(x_0) \\ &= \int_{X_1} \mathbf{1}_{A_1}\, d\mu_1(x_0) \\ &= \mu_1(x_0)(A_1), \end{aligned}$$

for all $x_0 \in X_0$ and $A_1 \in \mathcal{A}_1$, where $\pi_1 : X_1 \times X_2 \to X_1$ is the canonical projection. Hence

$$\mu_1 = \lambda x_0.((\mu_1 \otimes \mu_2)(x_0) \circ \pi_1^{-1}).$$

Thus μ_1 is the marginal probability kernel for $\mu_1 \otimes \mu_2$ with respect to X_1.

Related to the remarks in the preceding note, here is a result about distributions of random variables mapping into product spaces.

Proposition A.7.2. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ measurable spaces, $\mu_1 : \mathcal{P}(X_1)$ a probability measure, $\mu_2 : X_1 \to \mathcal{P}(X_2)$ a probability kernel, and $f : \Omega \to X_1 \times X_2$ a random variable such that $\mathcal{L}(f) = \mu_1 \otimes \mu_2$. Suppose that $f = (f_1, f_2)$, where $f_1 : \Omega \to X_1$ and $f_2 : \Omega \to X_2$. Then the following hold.*

1. $\mathcal{L}(f_1) = \mu_1$.
2. $\mathcal{L}(f_2) = \mu_1 \odot \mu_2 = \lambda A_2.\mathsf{E}(\lambda\omega.\mu_2(f_1(\omega))(A_2))$.
3. $\mathcal{L}(\mu_2 \circ f_1) = \mu_1 \circ \mu_2^{-1}$.

Proof. 1. Let $\pi_1 : X_1 \times X_2 \to X_1$ and $\pi_2 : X_1 \times X_2 \to X_2$ be the canonical projections. Then $\mathcal{L}(f_1) = \mathsf{P} \circ f_1^{-1} = (\mathsf{P} \circ f^{-1}) \circ \pi_1^{-1} = (\mu_1 \otimes \mu_2) \circ \pi_1^{-1} = \mu_1$.

2. Similarly, $\mathcal{L}(f_2) = \mathsf{P} \circ f_2^{-1} = (\mathsf{P} \circ f^{-1}) \circ \pi_2^{-1} = (\mu_1 \otimes \mu_2) \circ \pi_2^{-1} = \mu_1 \odot \mu_2$.

Now consider the random variable $\mu_2 \circ f_1 : \Omega \to \mathcal{P}(X_2)$. Recall that, by the definition of the σ-algebra on $\mathcal{P}(X_2)$, $\lambda\omega.\mu_2(f_1(\omega))(A_2) : \Omega \to \mathbb{R}$ is measurable, for all $A_2 \in \mathcal{A}_2$. Then

$$\begin{aligned}
&\mu_1 \odot \mu_2 \\
&= \lambda A_2. \int_{X_1} \lambda x_1.\mu_2(x_1)(A_2)\ d\mu_1 \\
&= \lambda A_2. \int_{X_1} \lambda x_1.\mu_2(x_1)(A_2)\ d(\mathsf{P} \circ f_1^{-1}) \\
&= \lambda A_2. \int_{\Omega} (\lambda x_1.\mu_2(x_1)(A_2) \circ f_1)\ d\mathsf{P} && \text{[Proposition A.2.14]} \\
&= \lambda A_2. \int_{\Omega} \lambda\omega.\mu_2(f_1(\omega))(A_2)\ d\mathsf{P} \\
&= \lambda A_2.\mathsf{E}(\lambda\omega.\mu_2(f_1(\omega))(A_2)).
\end{aligned}$$

3. The function $\mu_2 \circ f_1 : \Omega \to \mathcal{P}(X_2)$ is a random variable. Thus $\mathcal{L}(\mu_2 \circ f_1)$ is a distribution on the set of probability measures on X_2. Also $\mathcal{L}(\mu_2 \circ f_1) = \mathsf{P} \circ (\mu_2 \circ f_1)^{-1} = (\mathsf{P} \circ f_1^{-1}) \circ \mu_2^{-1} = \mu_1 \circ \mu_2^{-1}$. □

Suppose that

$$\begin{aligned}
&\mu_1 : X_0 \to \mathcal{P}(X_1) \\
&\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2) \\
&\mu_3 : X_0 \times X_1 \times X_2 \to \mathcal{P}(X_3) \\
&\quad\vdots \\
&\mu_{n-1} : X_0 \times X_1 \times \cdots \times X_{n-2} \to \mathcal{P}(X_{n-1}) \\
&\mu_n : X_0 \times X_1 \times \cdots \times X_{n-2} \times X_{n-1} \to \mathcal{P}(X_n)
\end{aligned}$$

are probability kernels. Then $\bigotimes_{i=1}^{n} \mu_i$ means $(\cdots(((\mu_1 \otimes \mu_2) \otimes \mu_3) \otimes \mu_4) \cdots \otimes \mu_n)$, and is well-defined, by Proposition A.7.1.

Proposition A.7.3. *Let $(X_i, \mathcal{A}_i)$ be a measurable space, for $i = 0, \ldots, n$, and $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ a probability kernel, for $i = 1, \ldots, n$. Then $\bigotimes_{i=1}^{n} \mu_i : X_0 \to \mathcal{P}(\prod_{i=1}^{n} X_i)$ is a probability kernel.*

Proof. By induction on n, using Proposition A.7.1. □

Taking X_0 to be a singleton set, the following corollary of Proposition A.7.1 is obtained.

Proposition A.7.4. *Let $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ be measurable spaces, $\mu_1 : \mathcal{P}(X_1)$ a probability measure, and $\mu_2 : X_1 \to \mathcal{P}(X_2)$ a probability kernel. Then $\mu_1 \otimes \mu_2 : \mathcal{P}(X_1 \times X_2)$. Furthermore, $\mu_1 \otimes \mu_2$ is uniquely defined by the property:*

$$(\mu_1 \otimes \mu_2)(A_1 \times A_2) = \int_{X_1} \mathbf{1}_{A_1} \, \lambda x_1.\mu_2(x_1)(A_2) \, d\mu_1,$$

for all $A \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$.

Note that, if $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ are probability kernels, then, for all $x_0 \in X_0$, $\mu_1(x_0) : \mathcal{P}(X_1)$, $\lambda x_1.\mu_2(x_0, x_1) : X_1 \to \mathcal{P}(X_2)$, and

$$(\mu_1 \otimes \mu_2)(x_0) = \mu_1(x_0) \otimes \lambda x_1.\mu_2(x_0, x_1).$$

Suppose that

$$\begin{aligned}
&\mu_1 : \mathcal{P}(X_1) \\
&\mu_2 : X_1 \to \mathcal{P}(X_2) \\
&\mu_3 : X_1 \times X_2 \to \mathcal{P}(X_3) \\
&\quad \vdots \\
&\mu_{n-1} : X_1 \times \cdots \times X_{n-2} \to \mathcal{P}(X_{n-1}) \\
&\mu_n : X_1 \times \cdots \times X_{n-2} \times X_{n-1} \to \mathcal{P}(X_n)
\end{aligned}$$

are probability kernels. Then $\bigotimes_{i=1}^{n} \mu_i$ means $(\cdots(((\mu_1 \otimes \mu_2) \otimes \mu_3) \otimes \mu_4) \cdots \otimes \mu_n)$, and is well-defined, by Proposition A.7.4.

Proposition A.7.5. *Let $(X_i, \mathcal{A}_i)$ be a measurable space and $\mu_i : \prod_{j=1}^{i-1} X_j \to \mathcal{P}(X_i)$ a probability kernel, for $i = 1, \ldots, n$. Then $\bigotimes_{i=1}^{n} \mu_i : \mathcal{P}(\prod_{i=1}^{n} X_i)$.*

Proof. By induction on n, using Proposition A.7.4. □

Proposition A.7.6. *Let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ be measurable spaces, and $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels. Suppose that $f : X_1 \times X_2 \to \mathbb{R}$ is a non-negative, measurable function. Then*

$$\int_{X_1 \times X_2} f \, d(\mu_1 \otimes \mu_2)(x_0) = \int_{X_1} \left(\lambda x_1. \int_{X_2} \lambda x_2.f(x_1, x_2) \, d\mu_2(x_0, x_1) \right) \, d\mu_1(x_0),$$

for all $x_0 \in X_0$.

Proof. See [87, Theorem 14.29]. □

The next result is used for computing integrals with respect to products of probability kernels.

Proposition A.7.7. *(Fubini theorem for probability kernels) Let $(X_i, \mathcal{A}_i)$ be a measurable space, for $i = 0, \ldots, n$, and $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ a probability kernel, for $i = 1, \ldots, n$. Suppose that $f : \prod_{i=1}^{n} X_i \to \mathbb{R}$ is a non-negative, measurable function. Then, for all $x_0 \in X_0$,*

$$\begin{aligned}&\int_{\prod_{i=1}^n X_i} f \; d(\bigotimes_{i=1}^{n} \mu_i)(x_0) \\ &= \int_{X_1} \Bigg(\lambda x_1. \int_{X_2} \Bigg(\lambda x_2. \; \ldots \; \int_{X_n} \lambda x_n. f(x_1, \ldots, x_n) \; d\mu_n(x_0, x_1, \ldots, x_{n-1}) \; \ldots \Bigg) \\ &\qquad\qquad d\mu_2(x_0, x_1) \Bigg) \; d\mu_1(x_0).\end{aligned}$$

Proof. By induction on n, using Proposition A.7.6. □

Fubini's theorem is of considerable theoretical importance, but it is not often much use for actually computing integrals. Here is an example of the kind of strong assumptions that need to be made in order to use Fubini's theorem in practice.

Example A.7.1. In a typical application of Proposition A.7.7, the function f is a utility function. To allow the integral to be computed, some assumptions are made about the form of f. To illustrate these, consider the 3-dimensional case, where

$$f : X_1 \times X_2 \times X_3 \to \mathbb{R}.$$

Suppose that f can be decomposed in an analogous way to the probability kernel, so that, for all $x_1 \in X_1$, $x_2 \in X_2$, and $x_3 \in X_3$,

$$f(x_1, x_2, x_3) = f_1(x_1).f_2(x_1)(x_2).f_3(x_1, x_2)(x_3),$$

where

$$\begin{aligned}&f_1 : X_1 \to \mathbb{R} \\ &f_2 : X_1 \to X_2 \to \mathbb{R} \\ &f_3 : X_1 \times X_2 \to X_3 \to \mathbb{R}.\end{aligned}$$

Then, for all $x_0 \in X_0$,

$$\begin{aligned}
&\int_{X_1 \times X_2 \times X_3} f \; d(\mu_1 \otimes \mu_2 \otimes \mu_3)(x_0) \\
&= \int_{X_1} \left(\lambda x_1. \int_{X_2} \left(\lambda x_2. \int_{X_3} \lambda x_3. f(x_1, x_2, x_3) \; d\mu_3(x_0, x_1, x_2) \right) \; d\mu_2(x_0, x_1) \right) \; d\mu_1(x_0) \\
&= \int_{X_1} \left(\lambda x_1. \int_{X_2} \left(\lambda x_2. \int_{X_3} \lambda x_3. f_1(x_1).f_2(x_1)(x_2).f_3(x_1, x_2)(x_3) \; d\mu_3(x_0, x_1, x_2) \right) \right. \\
&\qquad\qquad \left. d\mu_2(x_0, x_1) \right) \; d\mu_1(x_0) \\
&= \int_{X_1} \left(\lambda x_1. f_1(x_1) \int_{X_2} \left(\lambda x_2. f_2(x_1)(x_2) \int_{X_3} \lambda x_3. f_3(x_1, x_2)(x_3) \; d\mu_3(x_0, x_1, x_2) \right) \right. \\
&\qquad\qquad \left. d\mu_2(x_0, x_1) \right) \; d\mu_1(x_0)
\end{aligned}$$

Now suppose that μ_1, μ_2, and μ_3 are piecewise-constant. Suppose also that $f_2 : X_1 \to (X_2 \to \mathbb{R})$ is piecewise-constant (on X_1) and $f_3 : X_1 \times X_2 \to (X_3 \to \mathbb{R})$ is also piecewise-constant (on $X_1 \times X_2$). Then the integral, which is a function with domain X_0, can be further decomposed into a number of cases each of which is defined by the sum of a number of terms of the form

$$\left(\int_{X_1'} \lambda x_1. f_1(x_1) \; d\mu_1' \right) \left(\int_{X_2'} \lambda x_2. f_2'(x_2) \; d\mu_2' \right) \left(\int_{X_3'} \lambda x_3. f_3'(x_3) \; d\mu_3' \right),$$

where X_1' is a subset of X_1, X_2' is a subset of X_2, X_3' is a subset of X_3, μ_1' is one of the cases for μ_1, μ_2' is one of the cases for μ_2, μ_3' is one of the cases for μ_3, and f_2' is one of the cases for f_2 and f_3' is one of the cases for f_3. It is assumed that each of these integrals is sufficiently simple that it can now be computed (approximately).

It is generally more practical to compute integrals of the form $\int_{\prod_{i=1}^n X_i} f \; d \bigotimes_{i=1}^n \mu_i$ by Monte Carlo methods. Assume that it is possible to sample from the probability space $(X_i, \mathcal{A}_i, \mu_i(x_1, \ldots, x_{i-1}))$, for all $x_i \in X_i$ and $i = 1, \ldots, n$. Then it becomes possible to do ancestral sampling from the probability space $(\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i, \bigotimes_{i=1}^n \mu_i)$, and so standard Monte Carlo methods become available.

The next result shows how to integrate with respect to a fusion of probability kernels.

Proposition A.7.8. *Let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$, and $(X_2, \mathcal{A}_2)$ be measurable spaces, $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels, and $f : X_2 \to \mathbb{R}$ a measurable function that is integrable with respect to $(\mu_1 \odot \mu_2)(x_0)$, for all $x_0 \in X_0$. Then, for all $x_0 \in X_0$,*

$$\int_{X_2} f \; d(\mu_1 \odot \mu_2)(x_0) = \int_{X_1} \left(\lambda x_1. \int_{X_2} f \; d\mu_2(x_0, x_1) \right) \; d\mu_1(x_0).$$

Proof. Let f be an indicator function $\mathbf{1}_{A_2}$, where $A_2 \in \mathcal{A}_2$. Then, for all $x_0 \in X_0$,

$$\begin{aligned}
&\int_{X_2} \mathbf{1}_{A_2}\, d(\mu_1 \odot \mu_2)(x_0) \\
&= (\mu_1 \odot \mu_2)(x_0)(A_2) \\
&= \int_{X_1} \lambda x_1.\mu_2(x_0, x_1)(A_2)\, d\mu_1(x_0) \\
&= \int_{X_1} \left(\lambda x_1. \int_{X_2} \mathbf{1}_{A_2}\, d\mu_2(x_0, x_1)\right)\, d\mu_1(x_0).
\end{aligned}$$

Hence the result holds for measurable indicator functions. By linearity of the integral, it holds for simple functions. By Proposition A.2.1 and the monotone convergence theorem (Proposition A.2.2), it also holds for non-negative measurable functions f. Then, since f is integrable with respect to $(\mu_1 \odot \mu_2)(x_0, x_2)$, for all $x_0 \in X_0$ and $x_2 \in X_2$, the result follows. □

The next result will be needed when conditional independence properties hold in product spaces.

Proposition A.7.9. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, and $(X_i, \mathcal{A}_i)$ a measurable space, for $i = 1, \ldots, n$. Suppose that $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{i=1}^n X_i$ are measurable functions. Let $\mu : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$ be a regular conditional distribution of f given f_0, $S \subseteq \{1, \ldots, n\}$, and $\pi_S : \prod_{i=1}^n X_i \to \prod_{i \in S} X_i$ the canonical projection. Then $\lambda x_0.(\mu(x_0) \circ \pi_S^{-1}) : X_0 \to \mathcal{P}(\prod_{i \in S} X_i)$ is a regular conditional distribution of $\pi_S \circ f$ given f_0.*

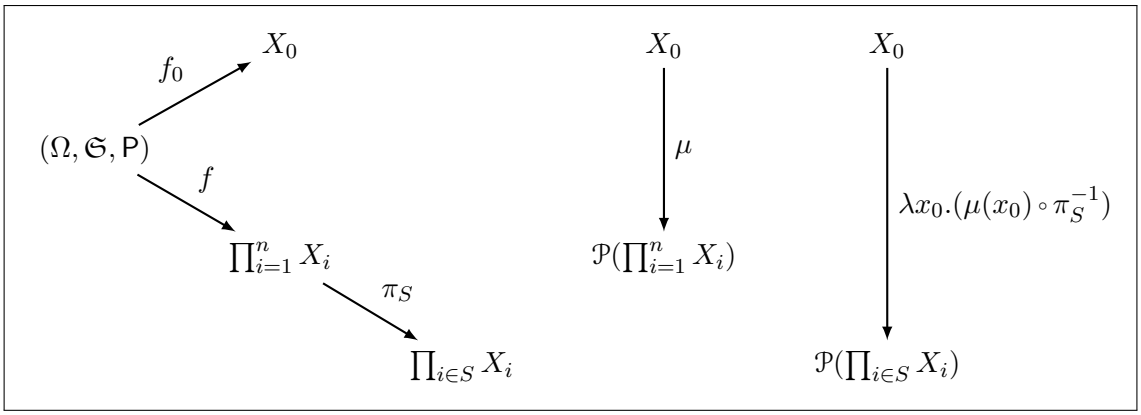

Figure A.19: Setting for Proposition A.7.9

Proof. Note first that Proposition A.2.13 shows that $\lambda x_0.(\mu(x_0) \circ \pi_S^{-1}) : X_0 \to \mathcal{P}(\prod_{i \in S} X_i)$ is well-defined.

Let $A \in \bigotimes_{i \in S} \mathcal{A}_i$. Then $\lambda x_0.(\mu(x_0) \circ \pi_S^{-1})(A) = \lambda x_0.\mu(x_0)(\pi_S^{-1}(A))$ is measurable, by Proposition A.2.4, since μ is a probability kernel and π_S is measurable. Hence, by Proposition A.2.4 again, $\lambda x_0.(\mu(x_0) \circ \pi_S^{-1})$ is a probability kernel.

Then, for all $A \in \bigotimes_{i \in S} \mathcal{A}_i$, almost surely,

$$\begin{aligned}
&\mathsf{P}((\pi_S \circ f)^{-1}(A) \mid f_0) \\
&= \mathsf{P}(f^{-1}(\pi_S^{-1}(A)) \mid f_0) \\
&= \lambda\omega.\mu(f_0(\omega))(\pi_S^{-1}(A)) && [\mu \text{ is a regular conditional distribution}] \\
&= \lambda\omega.\lambda x_0.(\mu(x_0) \circ \pi_S^{-1})(f_0(\omega))(A).
\end{aligned}$$

Hence $\lambda x_0.(\mu(x_0) \circ \pi_S^{-1})$ is a regular conditional distribution of $\pi_S \circ f$ given f_0. □

Next it is shown that a probability kernel which maps into the space of probability measures on a product space can be factored into a product of probability kernels which map into the space of probability measures on each factor of the product space.

Proposition A.7.10. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, and $(X_i, \mathcal{A}_i)$ a standard Borel space, for $i = 1, 2$. Suppose that $f_0 : \Omega \to X_0$ and $f : \Omega \to X_1 \times X_2$ are measurable. Let $\mu : X_0 \to \mathcal{P}(X_1 \times X_2)$ be a regular conditional distribution of f given f_0. Then there exist a regular conditional distribution $\mu_1 : X_0 \to \mathcal{P}(X_1)$ of f_1 given f_0 and a regular conditional distribution $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ of f_2 given (f_0, f_1) such that $\mu = \mu_1 \otimes \mu_2$ $\mathcal{L}(f_0)$-a.e.*

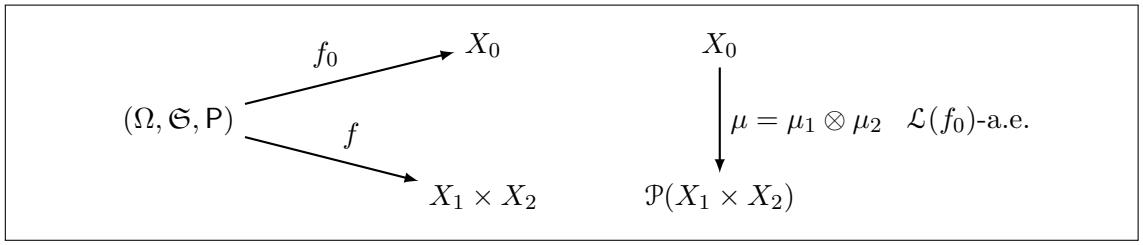

Figure A.20: Setting for Proposition A.7.10

Proof. Note that there exist measurable functions $f_1 : \Omega \to X_1$ and $f_2 : \Omega \to X_2$ such that $f = (f_1, f_2)$. Also, Proposition A.5.16 ensures that a probability kernel $\mu : X_0 \to \mathcal{P}(X_1 \times X_2)$ does exist that satisfies

$$\mathsf{P}(f^{-1}(A) \,|\, f_0) = \lambda\omega.\mu(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$.

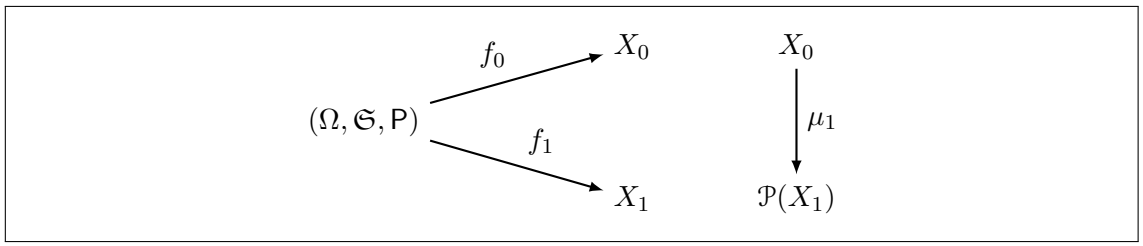

Figure A.21: Setting for Proposition A.7.10

By Proposition A.5.16, there is a probability kernel $\mu_1 : X_0 \to \mathcal{P}(X_1)$ such that

$$\mathsf{P}(f_1^{-1}(A_1) \,|\, f_0) = \lambda\omega.\mu_1(f_0(\omega))(A_1) \text{ a.s.},$$

for all $A_1 \in \mathcal{A}_1$.

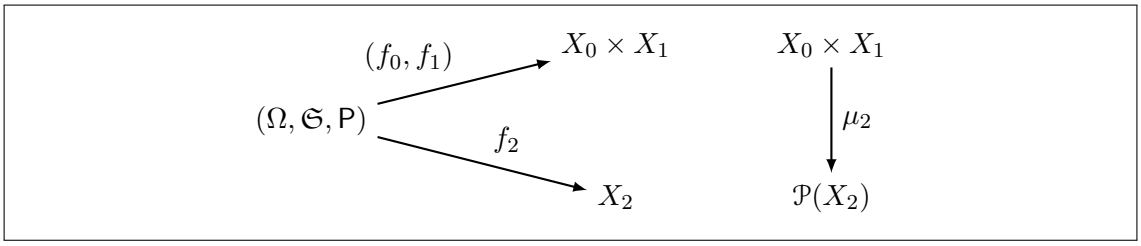

Figure A.22: Setting for Proposition A.7.10

Again, by Proposition A.5.16, there is a probability kernel $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ such that

$$\mathsf{P}(f_2^{-1}(A_2) \,|\, (f_0, f_1)) = \lambda\omega.\mu_2((f_0, f_1)(\omega))(A_2) \text{ a.s.},$$

for all $A_2 \in \mathcal{A}_2$.

Let

$$\mathcal{P} \triangleq \{A_1 \times A_2 \mid A_1 \in \mathcal{A}_1 \text{ and } A_2 \in \mathcal{A}_2\}$$

and

$$\mathcal{L} \triangleq \{A \in \mathcal{A}_1 \otimes \mathcal{A}_2 \mid \mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\mu_1 \otimes \mu_2)(f_0(\omega))(A) \text{ a.s.}\}.$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \mathcal{A}_1 \otimes \mathcal{A}_2$.

Suppose that $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(A_1 \times A_2) \mid f_0) \\
&= \mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \mathsf{E}(\mathbf{1}_{f_2^{-1}(A_2)} \mid (f_0, f_1)) \mid f_0) && \text{[Proposition A.5.10]} \\
&= \mathsf{E}(\mathbf{1}_{f_1^{-1}(A_1)}\, \lambda\omega.\mu_2((f_0, f_1)(\omega))(A_2) \mid f_0) \\
&= \lambda\omega. \int_{X_1} \mathbf{1}_{A_1}\, \lambda x_1.\mu_2(f_0(\omega), x_1)(A_2)\, d\mu_1(f_0(\omega)) && \text{[Proposition A.5.18]} \\
&= \lambda\omega.(\mu_1 \otimes \mu_2)(f_0(\omega))(A_1 \times A_2).
\end{aligned}$$

Thus $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, clearly $X_1 \times X_2 \in \mathcal{L}$. Second, suppose

that $(A_k)_{k\in\mathbb{N}}$ is a increasing sequence in $\mathcal{L}$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(\bigcup_{k\in\mathbb{N}} A_k) \mid f_0)\\
&= \mathsf{P}(\bigcup_{k\in\mathbb{N}} f^{-1}(A_k) \mid f_0)\\
&= \mathsf{E}(\lim_{k\to\infty} \mathbf{1}_{f^{-1}(A_k)} \mid f_0)\\
&= \lim_{k\to\infty} \mathsf{E}(\mathbf{1}_{f^{-1}(A_k)} \mid f_0) && \text{[Proposition A.5.8]}\\
&= \lim_{k\to\infty} \lambda\omega.(\mu_1\otimes\mu_2)(f_0(\omega))(A_k)\\
&= \lambda\omega.(\mu_1\otimes\mu_2)(f_0(\omega))(\bigcup_{k\in\mathbb{N}} A_k).
\end{aligned}$$

Thus $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{L}$. Third, suppose that $A, B \in \mathcal{L}$ and $A \subseteq B$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(B \setminus A) \mid f_0)\\
&= \mathsf{P}(f^{-1}(B) \setminus f^{-1}(A) \mid f_0)\\
&= \mathsf{E}(\mathbf{1}_{f^{-1}(B)} - \mathbf{1}_{f^{-1}(A)} \mid f_0) && [A \subseteq B]\\
&= \mathsf{P}(f^{-1}(B) \mid f_0) - \mathsf{P}(f^{-1}(A) \mid f_0) && \text{[Proposition A.5.6, Part 1]}\\
&= \lambda\omega.(\mu_1\otimes\mu_2)(f_0(\omega))(B) - \lambda\omega.(\mu_1\otimes\mu_2)(f_0(\omega))(A)\\
&= \lambda\omega.(\mu_1\otimes\mu_2)(f_0(\omega))(B \setminus A).
\end{aligned}$$

Thus $B \setminus A \in \mathcal{L}$. Hence it has been shown that $\mathcal{L}$ is a λ-system.

By Proposition A.1.2, $\sigma(\mathcal{P}) \subseteq \mathcal{L}$; that is, $\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\mu_1\otimes\mu_2)(f_0(\omega))(A)$ a.s., for all $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$. Thus, by the uniqueness part of Proposition A.5.16, it follows that $\mu = \mu_1 \otimes \mu_2$ $\mathcal{L}(f_0)$-a.e. □

As a corollary of Proposition A.7.10, the next result gives the factorization of a probability measure on a product space.

Proposition A.7.11. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ standard Borel spaces. Suppose that $f : \Omega \to X_1 \times X_2$ is measurable. Let μ be $\mathcal{L}(f)$. Then there exist a probability measure $\mu_1 : \mathcal{P}(X_1)$ and a regular conditional distribution $\mu_2 : X_1 \to \mathcal{P}(X_2)$ of f_2 given f_1 such that $\mu = \mu_1 \otimes \mu_2$.*

Proof. Let $X_0 \triangleq \{x_0\}$ be a singleton set with the σ-algebra $\{\{\}, \{x_0\}\}$. Define $f_0 : \Omega \to X_0$ by $f_0(\omega) = x_0$, for all $\omega \in \Omega$, and $\overline{\mu} : X_0 \to \mathcal{P}(X_1 \times X_2)$ by $\overline{\mu}(x_0) = \mu$. Then f_0 is measurable. Also $\overline{\mu}$ is a regular conditional distribution of f given f_0, since $\mathsf{P}(f^{-1}(A) \,|\, f_0) = \lambda\omega.\mathsf{P}(f^{-1}(A)) = \lambda\omega.\mu(A) = \lambda\omega.\overline{\mu}(f_0(\omega))(A)$ a.s. By Proposition A.7.10, there exist a regular conditional distribution $\overline{\mu}_1 : X_0 \to \mathcal{P}(X_1)$ of f_1 given f_0 and a regular conditional distribution $\overline{\mu}_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ of f_2 given (f_0, f_1) such that $\overline{\mu} = \overline{\mu}_1 \otimes \overline{\mu}_2$ $\mathcal{L}(f_0)$-a.e. Thus $\overline{\mu}(x_0) = \overline{\mu}_1(x_0) \otimes \lambda x_1.\overline{\mu}_2(x_0, x_1)$. Now $\overline{\mu}(x_0) = \mu$. Also define $\mu_1 \triangleq \overline{\mu}_1(x_0) \in \mathcal{P}(X_1)$ and $\mu_2 \triangleq \lambda x_1.\overline{\mu}_2(x_0, x_1) : X_1 \to \mathcal{P}(X_2)$. Note that μ_2 is a regular conditional distribution of f_2 given f_1, since $\mathsf{P}(f_2^{-1}(A_2) \,|\, f_1) = \mathsf{P}(f_2^{-1}(A_2) \,|\, (f_0, f_1)) = \lambda\omega.\overline{\mu}_2((f_0, f_1)(\omega))(A_2) = \lambda\omega.\mu_2(f_1(\omega))(A_2)$ a.s. Finally, $\mu = \mu_1 \otimes \mu_2$. □

Next is a result which states that if μ_i is a regular conditional distribution of f_i given $(f_0, \ldots, f_{i-1})$, for $i = 1, \ldots, n$, then $\bigotimes_{i=1}^n \mu_i$ is a regular conditional distribution of $(f_1, \ldots, f_n)$ given f_0.

Proposition A.7.12. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(X_i, \mathcal{A}_i)$ a measurable space, for $i = 0, \ldots, n$. Suppose that $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{i=1}^n X_i$ are measurable, and, for $i = 1, \ldots, n$, $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ is a probability kernel such that*

$$\mathsf{P}(f_i^{-1}(A_i) \mid (f_0, \ldots, f_{i-1})) = \lambda\omega.\mu_i((f_0, \ldots, f_{i-1})(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$. Then

$$\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^n \mu_i)(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$.

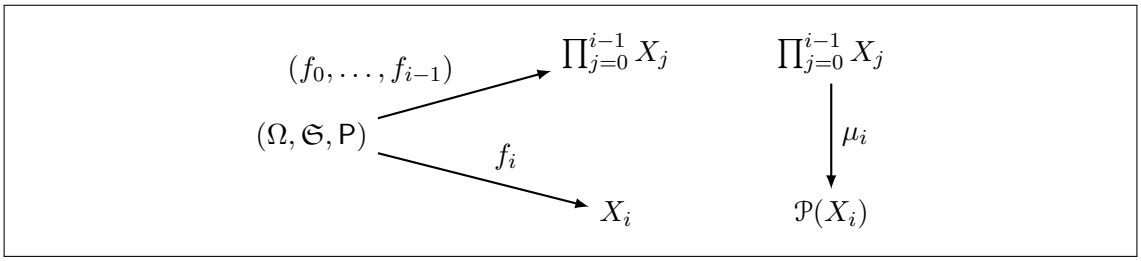

Figure A.23: Setting for Proposition A.7.12

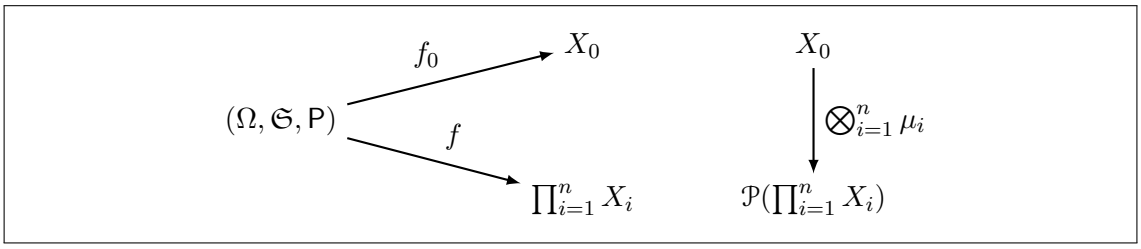

Figure A.24: Setting for Proposition A.7.12

Proof. Note that there exist measurable functions $f_i : \Omega \to X_i$, for $i = 1, \ldots, n$ such that $f = (f_1, \ldots, f_n)$.

The proof is by induction on n. For the base case, the result is obvious.

Assume now that the result holds for $n - 1$. Hence

$$\mathsf{P}((f_1, \ldots, f_{n-1})^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n-1} \mu_i)(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^{n-1} \mathcal{A}_i$.

Let

$$\mathcal{P} \triangleq \{\prod_{i=1}^{n} A_i \mid A_i \in \mathcal{A}_i, \text{ for } i = 1, \ldots, n\}$$

and

$$\mathcal{L} \triangleq \{A \in \bigotimes_{i=1}^{n} \mathcal{A}_i \mid \mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A) \text{ a.s.}\}.$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \bigotimes_{i=1}^{n} \mathcal{A}_i$.

Suppose that $A \triangleq \prod_{i=1}^{n-1} A_i \in \bigotimes_{i=1}^{n-1} \mathcal{A}_i$ and $A_n \in \mathcal{A}_n$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(A \times A_n) \mid f_0) \\
&= \mathsf{E}(\mathbf{1}_{(f_1,\ldots,f_{n-1})^{-1}(A)}\, \mathsf{E}(\mathbf{1}_{f_n^{-1}(A_n)} \mid (f_0, \ldots, f_{n-1})) \mid f_0) && \text{[Proposition A.5.10]} \\
&= \mathsf{E}(\mathbf{1}_{(f_1,\ldots,f_{n-1})^{-1}(A)}\, \lambda\omega.\mu_n((f_0, \ldots, f_{n-1})(\omega))(A_n) \mid f_0) \\
&= \lambda\omega. \int_{\prod_{i=1}^{n-1} X_i} \mathbf{1}_A\, \lambda(x_1, \ldots, x_{n-1}).\mu_n(f_0(\omega), x_1, \ldots, x_{n-1})(A_n)\, d(\bigotimes_{i=1}^{n-1} \mu_i)(f_0(\omega)) \\
&\qquad\qquad \text{[Proposition A.5.18 and induction hypothesis]} \\
&= \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A \times A_n).
\end{aligned}$$

Thus $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, clearly $\prod_{i=1}^{n} X_i \in \mathcal{L}$. Second, suppose that $(A_k)_{k\in\mathbb{N}}$ is a increasing sequence in $\mathcal{L}$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(\bigcup_{k\in\mathbb{N}} A_k) \mid f_0) \\
&= \mathsf{P}(\bigcup_{k\in\mathbb{N}} f^{-1}(A_k) \mid f_0) \\
&= \mathsf{E}(\lim_{k\to\infty} \mathbf{1}_{f^{-1}(A_k)} \mid f_0) \\
&= \lim_{k\to\infty} \mathsf{E}(\mathbf{1}_{f^{-1}(A_k)} \mid f_0) && \text{[Proposition A.5.8]} \\
&= \lim_{k\to\infty} \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A_k) \\
&= \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(\bigcup_{k\in\mathbb{N}} A_k).
\end{aligned}$$

Thus $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{L}$. Third, suppose that $A, B \in \mathcal{L}$ and $A \subseteq B$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(B \setminus A) \mid f_0) \\
&= \mathsf{P}(f^{-1}(B) \setminus f^{-1}(A) \mid f_0) \\
&= \mathsf{E}(\mathbf{1}_{f^{-1}(B)} - \mathbf{1}_{f^{-1}(A)} \mid f_0) && [A \subseteq B]
\end{aligned}$$

$$= \mathsf{P}(f^{-1}(B) \mid f_0) - \mathsf{P}(f^{-1}(A) \mid f_0) \qquad \text{[Proposition A.5.6, Part 1]}$$
$$= \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(B) - \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A)$$
$$= \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(B \setminus A).$$

Thus $B \setminus A \in \mathcal{L}$. Hence it has been shown that $\mathcal{L}$ is a λ-system.

By Proposition A.1.2, $\sigma(\mathcal{P}) \subseteq \mathcal{L}$; that is, $\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A)$ a.s., for all $A \in \bigotimes_{i=1}^{n} \mathcal{A}_i$. □

A corollary of Proposition A.7.12 is worth noting.

Proposition A.7.13. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(X_i, \mathcal{A}_i)$ a measurable space, for $i = 1, \ldots, n$. Suppose that $f : \Omega \to \prod_{i=1}^{n} X_i$ is measurable, and, for $i = 1, \ldots n$, $\mu_i : \prod_{j=1}^{i-1} X_j \to \mathcal{P}(X_i)$ is a probability kernel such that*

$$\mathsf{P}(f_i^{-1}(A_i) \mid (f_1, \ldots, f_{i-1})) = \lambda\omega.\mu_i((f_1, \ldots, f_{i-1})(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$. Then

$$\mathsf{P}(f^{-1}(A)) = (\bigotimes_{i=1}^{n} \mu_i)(A),$$

for all $A \in \bigotimes_{i=1}^{n} \mathcal{A}_i$.

Proof. Let X_0 be a singleton set $\{*\}$ and $f_0 : \Omega \to \{*\}$ the constant function. Note that $\sigma(f_0)$ is $\{\emptyset, \Omega\}$. Thus, by Part 8 (or Part 9) of Proposition A.5.6, $\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.\mathsf{P}(f^{-1}(A))$. The result now follows directly from Proposition A.7.12. □

Another way of stating the conclusion of Proposition A.7.13 is that $\mathcal{L}(f) = \bigotimes_{i=1}^{n} \mu_i$.

Here is the second version of Bayes theorem, this time for probability kernels.

Notation. For $A \subseteq X_1 \times X_2$, let $A^* \triangleq \{(x_2, x_1) \mid (x_1, x_2) \in A\}$.

Proposition A.7.14. *(Bayes theorem for probability kernels) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ standard Borel spaces, and $f_0 : \Omega \to X_0$, $f_1 : \Omega \to X_1$, and $f_2 : \Omega \to X_2$ random variables. Suppose that $\mu_1 : X_0 \to \mathcal{P}(X_1)$ is a regular conditional distribution of f_1 given f_0 and $\mu_{1,2} : X_0 \times X_1 \to \mathcal{P}(X_2)$ is a regular conditional distribution of f_2 given (f_0, f_1). Then there exist a regular conditional distribution $\mu_2 : X_0 \to \mathcal{P}(X_2)$ of f_2 given f_0 and a regular conditional distribution $\mu_{2,1} : X_0 \times X_2 \to \mathcal{P}(X_1)$ of f_1 given (f_0, f_2) such that, $\mathcal{L}(f_0)$-almost surely,*

$$\lambda x_0.(\mu_2 \otimes \mu_{2,1})(x_0)(A^*) = \lambda x_0.(\mu_1 \otimes \mu_{1,2})(x_0)(A), \quad \text{for all } A \in \mathcal{A}_1 \otimes \mathcal{A}_2.$$

Proof. By Proposition A.7.12, the probability kernel $\mu_1 \otimes \mu_{1,2} : X_0 \to \mathcal{P}(X_1 \times X_2)$ is a regular conditional distribution of (f_1, f_2) given f_0. Now consider $\mu : X_0 \to \mathcal{P}(X_2 \times X_1)$ defined by

$$\mu(x_0)(A^*) = (\mu_1 \otimes \mu_{1,2})(x_0)(A),$$

for all $x_0 \in X_0$ and $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$. Then μ is a probability kernel that is a conditional probability distribution of (f_2, f_1) given f_0. By Proposition A.7.10, there exist a regular conditional distribution $\mu_2 : X_0 \to \mathcal{P}(X_2)$ of f_2 given f_0 and a regular conditional distribution $\mu_{2,1} : X_0 \times X_2 \to \mathcal{P}(X_1)$ of f_1 given (f_0, f_2) such that $\mu = \mu_2 \otimes \mu_{2,1}$ $\mathcal{L}(f_0)$-a.e. Thus, $\mathcal{L}(f_0)$-almost surely,

$$\lambda x_0.(\mu_2 \otimes \mu_{2,1})(x_0)(A^*) = \lambda x_0.(\mu_1 \otimes \mu_{1,2})(x_0)(A), \quad \text{for all } A \in \mathcal{A}_1 \otimes \mathcal{A}_2.$$

□

Definition A.7.2. With Bayes theorem for probability kernels in the form

$$\lambda x_0.(\mu_2 \otimes \mu_{2,1})(x_0)(A^*) = \lambda x_0.(\mu_1 \otimes \mu_{1,2})(x_0)(A),$$

μ_1 is the *prior*, $\mu_{1,2}$ is the *likelihood*, and $\mu_{2,1}$ is the *posterior*.

Later, an approach for isolating the posterior $\mu_{2,1}$ is discussed. (See Proposition A.12.8.)

An analogous result to Proposition A.7.12 for the fusion of probability kernels will be needed.

Proposition A.7.15. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$, and $(X_2, \mathcal{A}_2)$, measurable spaces, and $f_0 : \Omega \to X_0$, $f_1 : \Omega \to X_1$, and $f_2 : \Omega \to X_2$ random variables. Suppose that $\mu_1 : X_0 \to \mathcal{P}(X_1)$ is a probability kernel such that*

$$\mathsf{P}(f_1^{-1}(A_1) \mid f_0) = \lambda\omega.\mu_1(f_0(\omega))(A_1) \text{ a.s., for all } A_1 \in \mathcal{A}_1,$$

and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ a probability kernel such that

$$\mathsf{P}(f_2^{-1}(A_2) \mid (f_0, f_1)) = \lambda\omega.\mu_2((f_0, f_1)(\omega))(A_2) \text{ a.s., for all } A_2 \in \mathcal{A}_2.$$

Then $\mu_1 \odot \mu_2 : X_0 \to \mathcal{P}(X_2)$ satisfies

$$\mathsf{P}(f_2^{-1}(A_2) \mid f_0) = \lambda\omega.(\mu_1 \odot \mu_2)(f_0(\omega))(A_2) \text{ a.s., for all } A_2 \in \mathcal{A}_2.$$

Proof. For all $A_2 \in \mathcal{A}_2$, almost surely,

$$\begin{aligned}
&\lambda\omega.(\mu_1 \odot \mu_2)(f_0(\omega))(A_2) \\
&= \lambda\omega.(\mu_1 \otimes \mu_2)(f_0(\omega))(X_1 \times A_2) \\
&= \mathsf{P}((f_1, f_2)^{-1}(X_1 \times A_2) \mid f_0) \qquad \text{[Proposition A.7.12]} \\
&= \mathsf{P}(f_2^{-1}(A_2) \mid f_0).
\end{aligned}$$

□

Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ measurable spaces, $f : \Omega \to X$ and $g : \Omega \to Y$ random variables, and $\mu : X \to \mathcal{P}(Y)$ a regular conditional distribution of g given f. Suppose that the definition of μ is not known, but is needed for some application. In such a case, μ may be approximated in a useful way as follows. Suppose that there exists a measurable space $(Z, \mathcal{C})$, a random variable $h : \Omega \to Z$, a regular conditional distribution $\mu_1 : X \to \mathcal{P}(Z)$ of h given f, and a regular conditional distribution

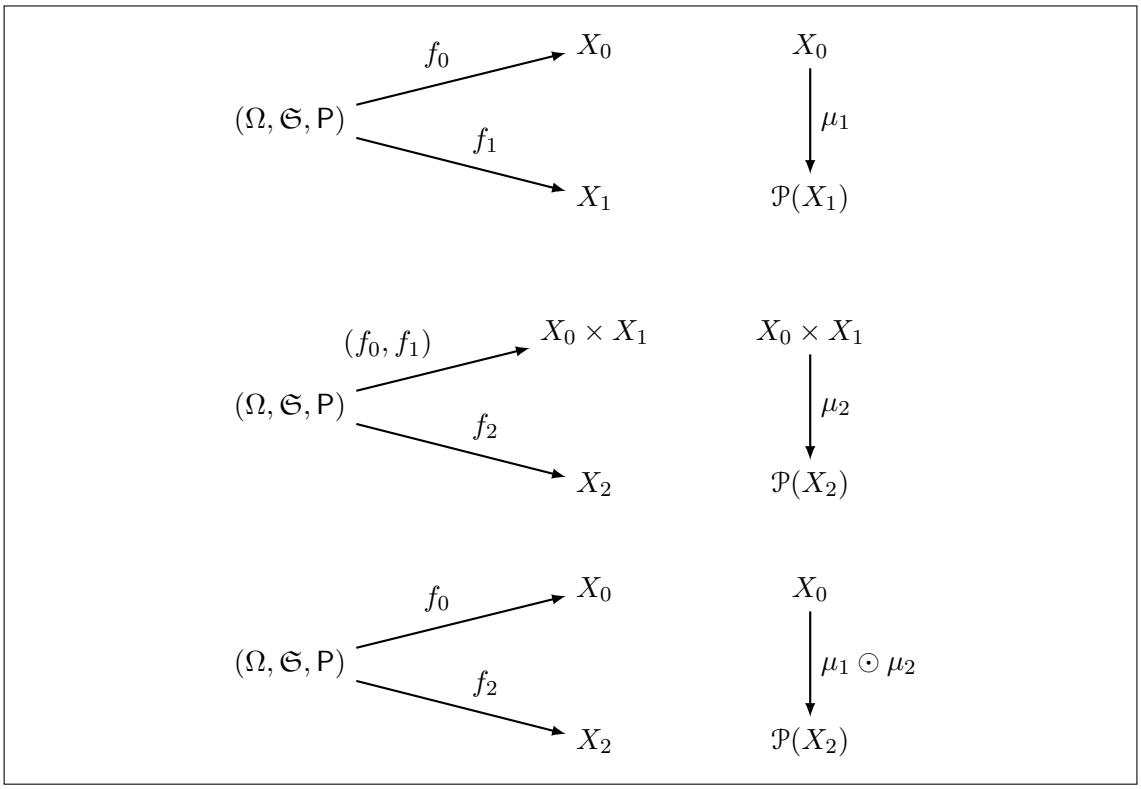

Figure A.25: Setting for Proposition A.7.15

$\mu_2 : Z \to \mathcal{P}(Y)$ of g given h, and that the definitions of μ_1 and μ_2 are known. (See Figure A.26.) Now, by Proposition A.2.9, $\lambda\gamma.(\gamma \odot \mu_2) : \mathcal{P}(Z) \to \mathcal{P}(Y)$ is a probability kernel. Hence $\lambda\gamma.(\gamma \odot \mu_2) \circ \mu_1 : X \to \mathcal{P}(Y)$ is a probability kernel that can be used to approximate μ. It would be better if $\mu = \lambda\gamma.(\gamma \odot \mu_2) \circ \mu_1$, but this is likely to be a rare occurrence. The next example gives a counterexample to show that, in general, $\mu \neq \lambda\gamma.(\gamma \odot \mu_2) \circ \mu_1$. The strategy is to select Z so that the definitions of μ_1 and μ_2 are known and the approximation of μ by $\lambda\gamma.(\gamma \odot \mu_2) \circ \mu_1$ is as close as necessary.

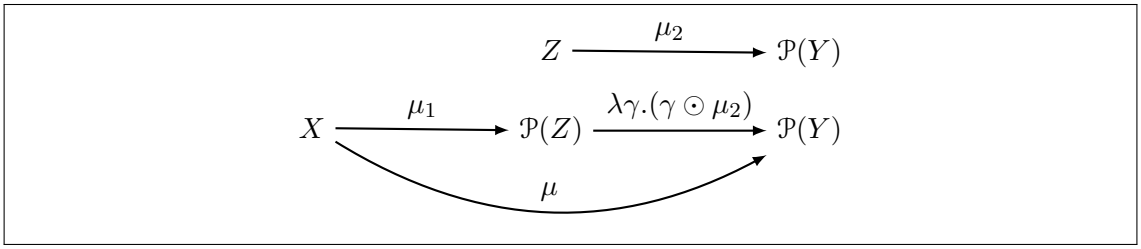

Figure A.26: Setting for Example A.7.2

Example A.7.2. Consider the setting of Example A.5.2. The probability kernel $\mu : X \to \mathcal{P}(Y)$ there is approximated as follows. In addition to the ingredients of Example A.5.2, suppose that $Z = \{e\}$, and $h : \Omega \to Z$ is the unique random variable with codomain Z. Suppose that $\mu_1 : X \to \mathcal{P}(Z)$ is defined by $\mu_1 = \lambda x.\epsilon$, where $\epsilon \in \mathcal{P}(Z)$ is the unique

probability measure on Z. Note that $\mu_1 : X \to \mathcal{P}(Z)$ is a regular conditional distribution of h gven f. Suppose also that $\mu_2 : Z \to \mathcal{P}(Y)$ is defined by $\mu_2(e) = \nu$, for the particular probability measure ν on Y such that μ_2 is a regular conditional distribution of g gven h.

Now consider $\lambda\gamma.(\gamma \odot \mu_2) \circ \mu_1 : X \to \mathcal{P}(Y)$. Then, for all $x \in X$,

$$\begin{aligned}
&(\lambda\gamma.(\gamma \odot \mu_2) \circ \mu_1)(x) \\
&= (\lambda\gamma.(\gamma \odot \mu_2))(\mu_1(x)) \\
&= (\lambda\gamma.(\gamma \odot \mu_2))(\epsilon) \\
&= \epsilon \odot \mu_2 \\
&= \lambda B. \int_Z \lambda z.\mu_2(z)(B)\, d\epsilon \\
&= \lambda B. \int_Z \lambda z.\nu(B)\, d\epsilon \\
&= \lambda B.\nu(B) \\
&= \nu.
\end{aligned}$$

Thus $\lambda\gamma.(\gamma \odot \mu_2) \circ \mu_1$ is a constant function and hence is not equal to μ.

The next result is the converse of Proposition A.7.12. It states that if $\bigotimes_{i=1}^n \mu_i$ is a regular conditional distribution of f given f_0, then μ_i is a regular conditional distribution of f_i given $(f_0, \ldots, f_{i-1})$, for $i = 1, \ldots, n$.

Proposition A.7.16. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(X_i, \mathcal{A}_i)$ a measurable space, for $i = 0, \ldots, n$. Suppose that $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{i=1}^n X_i$ are measurable, and $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ is a probability kernel, for $i = 1, \ldots, n$, such that*

$$\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$. Then, for $i = 1, \ldots n$,

$$\mathsf{P}(f_i^{-1}(A_i) \mid (f_0, \ldots, f_{i-1})) = \lambda\omega.\mu_i((f_0, \ldots, f_{i-1})(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$.

Proof. The proof is by induction on n. The base case is obvious.

For the inductive step, suppose that the result holds for n. Suppose now that $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{i=1}^{n+1} X_i$ are measurable, and $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ is a probability kernel, for $i = 1, \ldots, n+1$, such that

$$\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n+1} \mu_i)(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^{n+1} \mathcal{A}_i$.

First, it is shown that

$$\mathsf{P}((f_1, \ldots, f_n)^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^{n} \mathcal{A}_i$. For this, almost surely,

$$\begin{aligned}
&\mathsf{P}((f_1, \ldots, f_n)^{-1}(A) \mid f_0) \\
&= \mathsf{P}(f^{-1}(A \times X_{n+1}) \mid f_0) \\
&= \lambda\omega.((\bigotimes_{i=1}^{n} \mu_i) \otimes \mu_{n+1})(f_0(\omega))(A \times X_{n+1}) \\
&= \lambda\omega. \int_{\prod_{i=1}^{n} X_i} \mathbf{1}_A \; \lambda(x_1, \ldots, x_n).\mu_{n+1}(f_0(\omega), x_1, \ldots, x_n)(X_{n+1}) \; d(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega)) \\
&= \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A).
\end{aligned}$$

Hence, by the inductive hypothesis, for $i = 1, \ldots, n$,

$$\mathsf{P}(f_i^{-1}(A_i) \mid (f_0, \ldots, f_{i-1})) = \lambda\omega.\mu_i((f_0, \ldots, f_{i-1})(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$.

Finally, it is shown that

$$\mathsf{P}(f_{n+1}^{-1}(A_{n+1}) \mid (f_0, \ldots, f_n)) = \lambda\omega.\mu_{n+1}((f_0, \ldots, f_n)(\omega))(A_{n+1}) \text{ a.s.},$$

for all $A_{n+1} \in \mathcal{A}_{n+1}$. For this, it suffices to show that

$$\int_\Omega \mathbf{1}_B \; \lambda\omega.\mu_{n+1}((f_0, \ldots, f_n)(\omega))(A_{n+1}) \; d\mathsf{P} = \int_\Omega \mathbf{1}_B \; \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})} \; d\mathsf{P},$$

for all $B \in \sigma(f_0, \ldots, f_n)$.

Put

$$\mathcal{P} = \{\prod_{i=0}^{n} A_i) \mid A_i \in \mathcal{A}_i, \text{ for } i = 0, \ldots, n\}$$

and

$$\begin{aligned}
\mathcal{L} = \{A \in &\bigotimes_{i=0}^{n} \mathcal{A}_i \mid \\
&\int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A)} \; \lambda\omega.\mu_{n+1}((f_0, \ldots, f_n)(\omega))(A_{n+1}) \; d\mathsf{P} = \int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A)} \; \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})} \; d\mathsf{P}\}.
\end{aligned}$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \bigotimes_{i=0}^{n} \mathcal{A}_i$.

Suppose that $A_i \in \mathcal{A}_i$, for $i = 0, \ldots, n$. Then, almost surely,

$$
\begin{aligned}
&\int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A_0\times\cdots\times A_n)}\ \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})}\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{f_0^{-1}(A_0)}\ \mathbf{1}_{f^{-1}(A_1\times\cdots\times A_{n+1})}\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{f_0^{-1}(A_0)}\ \mathsf{P}(f^{-1}(A_1\times\cdots\times A_{n+1}) \mid f_0)\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{f_0^{-1}(A_0)}\ \lambda\omega.(\bigotimes_{i=1}^{n+1}\mu_i)(f_0(\omega))(A_1\times\cdots\times A_{n+1})\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{f_0^{-1}(A_0)}\ \left(\lambda\omega.\int_{\prod_{i=1}^n X_i} \mathbf{1}_{A_1\times\cdots\times A_n}\right. \\
&\qquad\qquad \left.\lambda(x_1,\ldots,x_n).\mu_{n+1}(f_0(\omega),x_1,\ldots,x_n)(A_{n+1})\ d(\bigotimes_{i=1}^{n}\mu_i)(f_0(\omega))\right)\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{f_0^{-1}(A_0)}\ \mathsf{E}(\lambda(x_0,\ldots,x_n).\mathbf{1}_{X_0\times A_1\times\cdots\times A_n}(x_0,\ldots,x_n)\ \mu_{n+1}(x_0,\ldots,x_n)(A_{n+1}) \\
&\qquad\qquad \circ(f_0,\ldots,f_n) \mid f_0)\ d\mathsf{P} \quad \text{[Proposition A.5.18]} \\
&= \int_\Omega \mathbf{1}_{f_0^{-1}(A_0)}\ \mathsf{E}(\mathbf{1}_{(f_1,\ldots,f_n)^{-1}(A_1\times\cdots\times A_n)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1}) \mid f_0)\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{f_0^{-1}(A_0)}\ \mathbf{1}_{(f_1,\ldots,f_n)^{-1}(A_1\times\cdots\times A_n)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A_0\times\cdots\times A_n)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P}.
\end{aligned}
$$

Thus $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, $\prod_{i=0}^n X_i \in \mathcal{L}$, since $\prod_{i=0}^n X_i \in \mathcal{P}$. Second, suppose that $(A_k)_{k\in\mathbb{N}}$ is a increasing sequence in $\mathcal{L}$. Then, almost surely,

$$
\begin{aligned}
&\int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(\bigcup_{k\in\mathbb{N}} A_k)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{\bigcup_{k\in\mathbb{N}}(f_0,\ldots,f_n)^{-1}(A_k)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} \\
&= \int_\Omega \lim_{k\to\infty} \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A_k)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} \\
&= \lim_{k\to\infty}\int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A_k)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} \quad \text{[Dominated conv.]} \\
&= \lim_{k\to\infty}\int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A_k)}\ \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})}\ d\mathsf{P} \\
&= \int_\Omega \lim_{k\to\infty} \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A_k)}\ \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})}\ d\mathsf{P} \qquad \text{[Dominated conv.]} \\
&= \int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(\bigcup_{k\in\mathbb{N}} A_k)}\ \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})}\ d\mathsf{P}.
\end{aligned}
$$

Thus $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{L}$. Third, suppose that $A, B \in \mathcal{L}$ and $A \subseteq B$. Then, almost surely,

$$\begin{aligned}
&\int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(B\setminus A)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} \\
&= \int_\Omega (\mathbf{1}_{(f_0,\ldots,f_n)^{-1}(B)} - \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A)})\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} \quad [A\subseteq B] \\
&= \int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(B)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} - \\
&\qquad\qquad \int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A)}\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(B)}\ \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})}\ d\mathsf{P} \ -\ \int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(A)}\ \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})}\ d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_{(f_0,\ldots,f_n)^{-1}(B\setminus A)}\ \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})}\ d\mathsf{P}.
\end{aligned}$$

Thus $B \setminus A \in \mathcal{L}$. Hence it has been shown that $\mathcal{L}$ is a λ-system.

By Proposition A.1.2, $\sigma(\mathcal{P}) \subseteq \mathcal{L}$; that is

$$\int_\Omega \mathbf{1}_B\ \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1})\ d\mathsf{P} = \int_\Omega \mathbf{1}_B\ \mathbf{1}_{f_{n+1}^{-1}(A_{n+1})}\ d\mathsf{P},$$

for all $B \in \sigma(f_0,\ldots,f_n)$. Hence

$$\mathsf{P}(f_{n+1}^{-1}(A_{n+1}) \mid (f_0,\ldots,f_n)) = \lambda\omega.\mu_{n+1}((f_0,\ldots,f_n)(\omega))(A_{n+1}) \text{ a.s.},$$

for all $A_{n+1} \in \mathcal{A}_{n+1}$. □

A corollary of Proposition A.7.16 will be useful. It shows that each μ_j is a regular conditional distribution with respect to the probability measure $\bigotimes_{j=1}^n \mu_j$. Here, for all $i = 1,\ldots,n$, $\pi_i : \prod_{j=1}^n X_j \to X_i$ and $\pi_{1,\ldots,i-1} : \prod_{j=1}^n X_j \to \prod_{j=1}^{i-1} X_j$ are the canonical projections.

Proposition A.7.17. *Let $(X_i, \mathcal{A}_i)$ be a measurable space and $\mu_i : \prod_{j=1}^{i-1} X_j \to \mathcal{P}(X_i)$ a probability kernel, for $i = 1,\ldots,n$. Then, for all $i = 1,\ldots,n$ and $C_i \in \mathcal{A}_i$,*

$$(\bigotimes_{j=1}^n \mu_j)(\pi_i^{-1}(C_i) \mid \pi_{1,\ldots,i-1}) = \lambda x.\mu_i(\pi_{1,\ldots,i-1}(x))(C_i)\ \bigotimes_{j=1}^n \mu_j\text{-a.e.}$$

Proof. In Proposition A.7.16, let $\Omega \triangleq \prod_{j=1}^n X_j$, $\mathfrak{S} \triangleq \bigotimes_{j=1}^n \mathcal{A}_j$, and $\mathsf{P} \triangleq \bigotimes_{j=1}^n \mu_j$. Also let X_0 be a singleton set, f_0 the constant function onto X_0, and f the identity function. Then it is clear that $\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^n \mu_i)(f_0(\omega))(A)$ a.s., for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$. Hence Proposition A.7.16 applies. □

The next result is concerned with the common situation where a regular conditional distribution defined on a product space does not depend on all its arguments.

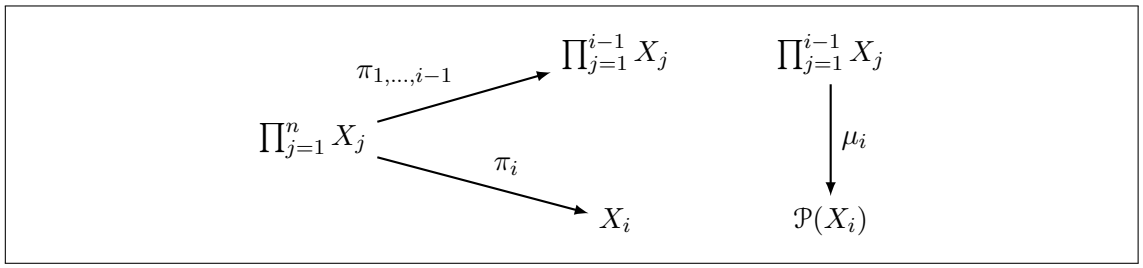

Figure A.27: Setting for Proposition A.7.17

Proposition A.7.18. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_i, \mathcal{A}_i)$ a measurable space and $f_i : \Omega \to X_i$ a measurable function, for $i = 1, \ldots, n$, $(Y, \mathcal{B})$ a measurable space, $g : \Omega \to Y$ a measurable function, and $\mu : \prod_{i=1}^n X_i \to \mathcal{P}(Y)$ a probability kernel such that*

$$\mathsf{P}(g^{-1}(B) \mid (f_1, \ldots, f_n)) = \lambda\omega.\mu((f_1, \ldots, f_n)(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{B}$. Suppose there exists a subset $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ and a probability kernel $\widehat{\mu} : \prod_{l=1}^k X_{i_l} \to \mathcal{P}(Y)$ such that $\mu = \lambda(x_1, \ldots, x_n).\widehat{\mu}(x_{i_1}, \ldots, x_{i_k})$. Then

$$\mathsf{P}(g^{-1}(B) \mid (f_{i_1}, \ldots, f_{i_k})) = \lambda\omega.\widehat{\mu}((f_{i_1}, \ldots, f_{i_k})(\omega))(B) \text{ a.s.},$$

for all $B \in \mathcal{B}$.

Proof. Note that $\sigma((f_{i_1}, \ldots, f_{i_k})) \subseteq \sigma((f_1, \ldots, f_n))$. Also, $\lambda\omega.\widehat{\mu}((f_{i_1}, \ldots, f_{i_k})(\omega))(B) : \Omega \to \mathbb{R}$ is $\sigma((f_{i_1}, \ldots, f_{i_k}))$-measurable, for all $B \in \mathcal{B}$. Then, for all $B \in \mathcal{B}$ and $C \in \sigma((f_{i_1}, \ldots, f_{i_k}))$,

$$\begin{aligned}
&\int_\Omega \mathbf{1}_C \, \mathbf{1}_{g^{-1}(B)} \, d\mathsf{P} \\
&= \int_\Omega \mathbf{1}_C \, \lambda\omega.\mu((f_1, \ldots, f_n)(\omega))(B) \, d\mathsf{P} \qquad [\mu \text{ is a regular conditional distribution}] \\
&= \int_\Omega \mathbf{1}_C \, \lambda\omega.\widehat{\mu}((f_{i_1}, \ldots, f_{i_k})(\omega))(B) \, d\mathsf{P}.
\end{aligned}$$

Hence $\widehat{\mu}$ is a regular conditional distribution of g given $(f_{i_1}, \ldots, f_{i_k})$. □

Here is the extension of Proposition A.7.10 to the general case.

Proposition A.7.19. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, and $(X_i, \mathcal{A}_i)$ a standard Borel space, for $i = 1, \ldots, n$. Suppose that $f_0 : \Omega \to X_0$ and $f_i : \Omega \to X_i$, for $i = 1, \ldots, n$, are measurable. Let $\mu : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$ be a regular conditional distribution of $(f_1, \ldots, f_n)$ given f_0. Then there exists a probability kernel $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ that is a regular conditional distribution of f_i given $(f_0, \ldots, f_{i-1})$, for $i = 1, \ldots, n$, such that $\mu = \bigotimes_{i=1}^n \mu_i$ $\mathcal{L}(f_0)$-a.e.*

Proof. By Proposition A.5.16, the probability kernel $\mu : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$ exists and satisfies

$$\mathsf{P}((f_1, \ldots, f_n)^{-1}(A) \,|\, f_0) = \lambda\omega.\mu(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$.

Also, by Proposition A.5.16, for $i = 1, \ldots n$, there is a probability kernel $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ such that

$$\mathsf{P}(f_i^{-1}(A_i) \mid (f_0, \ldots, f_{i-1})) = \lambda\omega.\mu_i((f_0, \ldots, f_{i-1})(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$.

By Proposition A.7.12, $\mathsf{P}((f_1, \ldots, f_n)^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^n \mu_i)(f_0(\omega))(A)$ a.s., for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$. Thus, by the uniqueness part of Proposition A.5.16, it follows that $\mu = \bigotimes_{i=1}^n \mu_i$ $\mathcal{L}(f_0)$-a.e. □

As a corollary of Proposition A.7.19, the next result gives the factorization of a probability measure on a product space.

Proposition A.7.20. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(X_i, \mathcal{A}_i)$ a standard Borel space, for $i = 1, \ldots, n$. Suppose that $f : \Omega \to \prod_{i=1}^n X_i$ is measurable. Let μ be $\mathcal{L}(f)$. Then there exist regular conditional distributions $\mu_i : \prod_{j=1}^{i-1} X_j \to \mathcal{P}(X_i)$ of f_i given $(f_1, \ldots, f_{i-1})$, for $i = 1, \ldots, n$, such that $\mu = \bigotimes_{i=1}^n \mu_i$.*

Proof. Let $X_0 \triangleq \{x_0\}$ be a singleton set with the σ-algebra $\{\{\}, \{x_0\}\}$. Define $f_0 : \Omega \to X_0$ by $f_0(\omega) = x_0$, for all $\omega \in \Omega$, and $\overline{\mu} : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$ by $\overline{\mu}(x_0) = \mu$. Then f_0 is measurable. Also $\overline{\mu}$ is a regular conditional distribution of $(f_1, \ldots, f_n)$ given f_0, since $\mathsf{P}((f_1, \ldots, f_n)^{-1}(A) \,|\, f_0) = \lambda\omega.\mathsf{P}((f_1, \ldots, f_n)^{-1}(A)) = \lambda\omega.\mu(A) = \lambda\omega.\overline{\mu}(f_0(\omega))(A)$ a.s. By Proposition A.7.19, there exists a probability kernel $\overline{\mu}_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ that is a regular conditional distribution of f_i given $(f_0, \ldots, f_{i-1})$, for $i = 1, \ldots, n$, such that $\overline{\mu} = \bigotimes_{i=1}^n \overline{\mu}_i$ $\mathcal{L}(f_0)$-a.e. Thus $\overline{\mu}(x_0) = \bigotimes_{i=1}^n \lambda(x_1, \ldots, x_{i-1}).\overline{\mu}_i(x_0, \ldots, x_{i-1})$. Now $\overline{\mu}(x_0) = \mu$. Also, for $i = 1, \ldots, n$, define $\mu_i : \prod_{j=1}^{i-1} X_j \to \mathcal{P}(X_i)$ by $\mu_i = \lambda(x_1, \ldots, x_{i-1}).\overline{\mu}_i(x_0, \ldots, x_{i-1})$. Clearly, each μ_i is a probability kernel. Also each μ_i is a regular conditional distribution of f_i given $(f_1, \ldots, f_{i-1})$, since $\mathsf{P}(f_i^{-1}(A_i) \,|\, (f_1, \ldots, f_{i-1})) = \mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, f_1, \ldots, f_{i-1})) = \lambda\omega.\overline{\mu}_i((f_0, f_1, \ldots, f_{i-1})(\omega))(A_i) = \lambda\omega.\mu_i((f_1, \ldots, f_{i-1})(\omega))(A_i)$ a.s. Finally, $\mu = \bigotimes_{i=1}^n \mu_i$. □

Note. In Proposition A.7.20, one can let $\Omega \triangleq \prod_{i=1}^n X_i$, $\mathsf{P} \triangleq \bigotimes_{i=1}^n \mu_i$, and let f be the identity mapping. Then $\mu = \mathcal{L}(f) = \bigotimes_{i=1}^n \mu_i$. Thus Proposition A.7.20 implies that every probability measure on a product of standard Borel spaces can be factorized.

Proposition A.7.21. *Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$ and $(Z_i, \mathcal{C}_i)$, for $i = 1, \ldots, n$, be measurable spaces and $\mu_i : X \times Y \times \prod_{j=1}^{i-1} Z_j \to \mathcal{P}(Z_i)$, for $i = 1, \ldots, n$, probability kernels. Then, for all $x \in X$,*

$$\lambda y.(\bigotimes_{i=1}^n \mu_i)(x, y) = \bigotimes_{i=1}^n \lambda(y, z_1, \ldots, z_{i-1}).\mu_i(x, y, z_1, \ldots, z_{i-1}).$$

Proof. Note that, by Proposition A.2.5, $\lambda y.(\bigotimes_{i=1}^n \mu_i)(x, y) : Y \to \mathcal{P}(\prod_{i=1}^n Z_i)$ and, for $i = 1, \ldots, n$, $\lambda(y, z_1, \ldots, z_{i-1}).\mu_i(x, y, z_1, \ldots, z_{i-1}) : Y \times \prod_{j=1}^{i-1} Z_j \to \mathcal{P}(Z_i)$ are probability kernels.

The proof is by induction on n. When $n = 1$, the result is obvious.

For the inductive step, assume that the result holds for n. Then, for all $x \in X$,

$$
\begin{aligned}
&\lambda y.(\bigotimes_{i=1}^{n+1} \mu_i)(x,y) \\
&= \lambda y.((\bigotimes_{i=1}^{n} \mu_i) \otimes \mu_{n+1}))(x,y) \\
&= \lambda y.\lambda C. \int_{\prod_{i=1}^{n} Z_i} (\lambda(z_1,\ldots,z_n). \\
&\qquad \int_{Z_{n+1}} \lambda z_{n+1}.\mathbf{1}_C(z_1,\ldots,z_{n+1})\; d\mu_{n+1}(x,y,z_1,\ldots,z_n)\Bigg)\; d(\bigotimes_{i=1}^{n} \mu_i)(x,y) \\
&= \lambda y.\lambda C. \int_{\prod_{i=1}^{n} Z_i} (\lambda(z_1,\ldots,z_n). \\
&\quad \int_{Z_{n+1}} \lambda z_{n+1}.\mathbf{1}_C(z_1,\ldots,z_{n+1})\; d\lambda(y,z_1,\ldots,z_n).\mu_{n+1}(x,y,z_1,\ldots,z_n)(y,z_1,\ldots,z_n)\Bigg) \\
&\qquad\qquad d\lambda y.(\bigotimes_{i=1}^{n} \mu_i)(x,y)(y) \\
&= \lambda y.\lambda C. \int_{\prod_{i=1}^{n} Z_i} (\lambda(z_1,\ldots,z_n). \\
&\quad \int_{Z_{n+1}} \lambda z_{n+1}.\mathbf{1}_C(z_1,\ldots,z_{n+1})\; d\lambda(y,z_1,\ldots,z_n).\mu_{n+1}(x,y,z_1,\ldots,z_n)(y,z_1,\ldots,z_n)\Bigg) \\
&\qquad\qquad d\bigotimes_{i=1}^{n} \lambda(y,z_1,\ldots,z_{i-1}).\mu_i(x,y,z_1,\ldots,z_{i-1})(y) \\
&\qquad\qquad\qquad \text{[Induction hypothesis]} \\
&= \bigotimes_{i=1}^{n+1} \lambda(y,z_1,\ldots,z_{i-1}).\mu_i(x,y,z_1,\ldots,z_{i-1}).
\end{aligned}
$$

Hence the result. □

The standard setting for deconstructing a probability kernel whose codomain is probability measures on a product space is given by Proposition A.7.19: the probability kernels $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$, for $i = 1,\ldots,n$, are needed. However, in practical situations, there are usually plentiful conditional independencies that can be exploited. These manifest themselves by probability kernels depending on only some (but not all) of the factors of $\prod_{j=0}^{i-1} X_j$. The following result extends Proposition A.7.19 by considering conditional independencies.

Notation. For $i = 1,\ldots,n$, suppose that $par(i) \triangleq \{i_1,\ldots,i_m\} \subseteq \{1,\ldots,i-1\}$, where $i_1 < \cdots < i_m$. Let $f_i : X \to Y_i$, for $i = 1,\ldots,n$. Then $f_{par(i)}$ denotes $(f_{i_1},\ldots,f_{i_m})$. Similarly, $x_{par(i)}$ denotes $(x_{i_1},\ldots,x_{i_m})$.

Proposition A.7.22. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, $(X_i, \mathcal{A}_i)$ a standard Borel space, for $i = 1, \ldots, n$, and $f_0 : \Omega \to X_0$ and $f_i : \Omega \to X_i$ measurable functions, for $i = 1, \ldots, n$. Suppose there is a dependency graph with vertices $0, \ldots, n$, where vertex i is labelled by $\sigma(f_i)$, for $i = 0, \ldots, n$, such that $0, \ldots, n$ is a topological order of the vertices and*

$$\sigma(f_i) \perp\!\!\!\perp_{\sigma(f_0, f_{par(i)})} \sigma(f_0, \ldots, f_{i-1}),$$

for $i = 1, \ldots, n$. Let $\mu : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$ be a regular conditional distribution of $(f_1, \ldots, f_n)$ given f_0. Then there exists a probability kernel $\mu_i : X_0 \times \prod_{j \in par(i)} X_j \to \mathcal{P}(X_i)$ that is a regular conditional distribution of f_i given $(f_0, f_{par(i)})$, for $i = 1, \ldots, n$, such that

$$\mu = \bigotimes_{i=1}^{n} \lambda(x_0, \ldots, x_{i-1}).\mu_i(x_0, x_{par(i)}) \ \ \mathcal{L}(f_0)\text{-a.e.}$$

Furthermore, $\lambda(x_0, \ldots, x_{i-1}).\mu_i(x_0, x_{par(i)})$ is regular conditional distribution of f_i given $(f_0, \ldots, f_{i-1})$, for $i = 1, \ldots, n$.

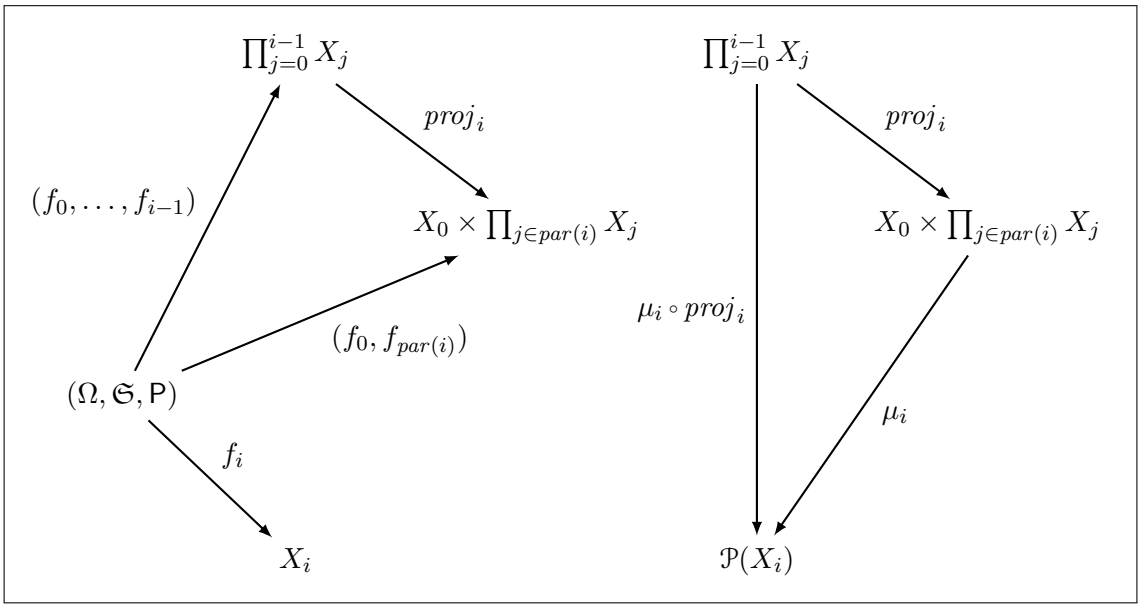

Figure A.28: In the figure, $proj_i$ denotes $\lambda(x_0, \ldots, x_{i-1}).(x_0, x_{par(i)})$

Proof. Since, for $i = 1, \ldots, n$, $\sigma(f_i) \perp\!\!\!\perp_{\sigma(f_0, f_{par(i)})} \sigma(f_0, \ldots, f_{i-1})$, it follows from Proposition A.6.1 that, for $i = 1, \ldots, n$,

$$\mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, \ldots, f_{i-1})) = \mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, f_{par(i)})) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$.

By Proposition A.5.16, for $i = 1, \ldots, n$, there exists a probability kernel $\mu_i : X_0 \times \prod_{j \in par(i)} X_j \to \mathcal{P}(X_i)$ such that

$$\mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, f_{par(i)})) = \lambda\omega.\mu_i((f_0, f_{par(i)})(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$. Hence, for $i = 1, \ldots, n$, and for all $A_i \in \mathcal{A}_i$, almost surely,

$$\begin{aligned}
&\mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, \ldots, f_{i-1})) \\
&= \mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, f_{par(i)})) \\
&= \lambda\omega.\mu_i((f_0, f_{par(i)})(\omega))(A_i) \\
&= \lambda\omega.\lambda(x_0, \ldots, x_{i-1}).\mu_i(x_0, x_{par(i)})((f_0, \ldots, f_{i-1})(\omega))(A_i).
\end{aligned}$$

Thus $\lambda(x_0, \ldots, x_{i-1}).\mu_i(x_0, x_{par(i)})$ is regular conditional distribution of f_i given $(f_0, \ldots, f_{i-1})$. By Proposition A.7.12,

$$\mathsf{P}((f_1, \ldots, f_n)^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n} \lambda(x_0, \ldots, x_{i-1}).\mu_i(x_0, x_{par(i)}))(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^{n} \mathcal{A}_i$. It follows that $\mu = \bigotimes_{i=1}^{n} \lambda(x_0, \ldots, x_{i-1}).\mu_i(x_0, x_{par(i)})$ $\mathcal{L}(f_0)$-a.e., by the uniqueness part of Proposition A.5.16. □

The next result is a corollary of Proposition A.7.22. It provides the theoretical basis for the practical success of Bayesian networks.

Proposition A.7.23. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, and $(X_i, \mathcal{A}_i)$ a standard Borel space and $f_i : \Omega \to X_i$ measurable functions, for $i = 1, \ldots, n$. Suppose there is a Markov dependency graph with vertices $1, \ldots, n$, where vertex i is labelled by $\sigma(f_i)$, for $i = 1, \ldots, n$, such that $1, \ldots, n$ is a topological order of the vertices. Let $\mu = \mathcal{L}(f)$, where $f \triangleq (f_1, \ldots, f_n)$. Then there exists a probability kernel $\mu_i : \prod_{j \in par(i)} X_j \to \mathcal{P}(X_i)$ that is a regular conditional distribution of f_i given $f_{par(i)}$, for $i = 1, \ldots, n$, such that*

$$\mu = \bigotimes_{i=1}^{n} \lambda(x_1, \ldots, x_{i-1}).\mu_i(x_{par(i)}).$$

Furthermore, $\lambda(x_1, \ldots, x_{i-1}).\mu_i(x_{par(i)})$ is regular conditional distribution of f_i given $(f_1, \ldots, f_{i-1})$, for $i = 1, \ldots, n$.

Proof. The fact that the dependency graph is Markov means precisely that

$$\sigma(f_i) \underset{\sigma(f_{par(i)})}{\perp\!\!\!\perp} \sigma(f_1, \ldots, f_{i-1}),$$

for $i = 1, \ldots, n$.

In Proposition A.7.22, let X_0 be a singleton set. Then $\mathcal{L}(f)$, which is $\mathsf{P} \circ f^{-1}$, can be identified with a regular conditional distribution of f given f_0. Hence the result follows directly from Proposition A.7.22. □

Note that $\lambda(x_1, \ldots, x_{i-1}).\mu_i(x_{par(i)}) = \mu_i \circ \lambda(x_1, \ldots, x_{i-1}).x_{par(i)}$, for $i = 1, \ldots, n$. In effect, each probability kernel $\lambda(x_1, \ldots, x_{i-1}).\mu_i(x_{par(i)})$ is factored through the projection $\lambda(x_1, \ldots, x_{i-1}).x_{par(i)}$.

In the context of random variables mapping into product spaces, further (implicit) conditional independencies can be found using Propositions A.6.4, A.6.5, and A.6.7.

The next result is used for computing integrals with respect to a product probability measure in the case of conditional independencies.

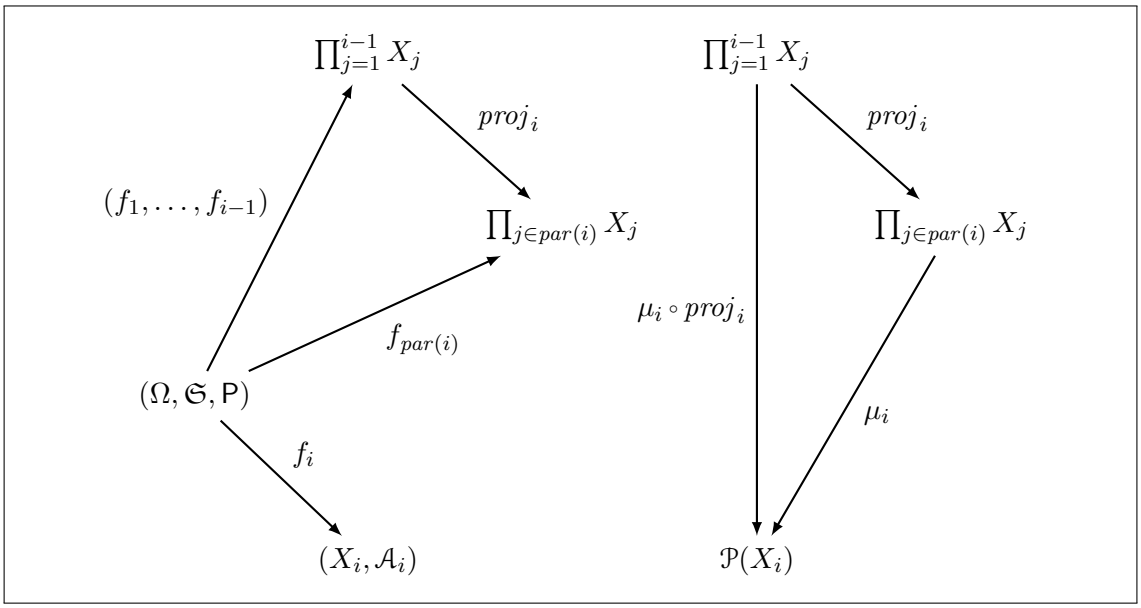

Figure A.29: In the figure, $proj_i$ denotes $\lambda(x_1, \ldots, x_{i-1}).x_{par(i)}$

Proposition A.7.24. *(Generalized Fubini theorem for probability kernels) Let $(X_i, \mathcal{A}_i)$ be a measurable space, for $i = 0, \ldots, n$. Suppose that there is a directed acyclic graph with vertices having labels from the set $\{1, \ldots, n\}$ such that $1, \ldots, n$ is a topological order of the vertices. For $i = 1, \ldots, n$, let $\mu_i : \prod_{j \in par(i)} X_j \to \mathcal{P}(X_i)$ be probability kernels. Suppose that $f : \prod_{i=1}^n X_i \to \mathbb{R}$ is a non-negative measurable function. Then*

$$\begin{aligned}
&\int_{\prod_{i=1}^n X_i} f \; d(\bigotimes_{i=1}^{n} \lambda(x_1, \ldots, x_{i-1}).\mu_i(x_{par(i)}))(x_0) \\
&= \int_{X_1} \left(\lambda x_1. \int_{X_2} \left(\lambda x_2. \; \ldots \; \int_{X_n} \lambda x_n. f(x_1, \ldots, x_n) \; d\mu_n(x_0, x_{par(n)}) \ldots \right) \right. \\
&\qquad \left. d\mu_2(x_0, x_{par(2)}) \right) \; d\mu_1(x_0).
\end{aligned}$$

Proof. The result follows directly from Proposition A.7.7. □

Example A.7.3. Let $(X_i, \mathcal{A}_i)$ be a measurable space, for $i = 1, \ldots, 5$. Consider the directed acyclic graph in Figure A.13. Suppose that

$\mu_1 : \mathcal{P}(X_1)$
$\mu_2 : \mathcal{P}(X_2)$
$\mu_3 : X_1 \times X_2 \to \mathcal{P}(X_3)$
$\mu_4 : X_3 \to \mathcal{P}(X_4)$
$\mu_5 : X_3 \to \mathcal{P}(X_5)$.

Thus we can form the product probability space

$$(\prod_{i=1}^{5} X_i, \bigotimes_{i=1}^{5} \mathcal{A}_i, \bigotimes_{i=1}^{5} \lambda(x_1, \ldots, x_{i-1}).\mu_i(x_{par(i)})).$$

If $f : \prod_{i=1}^{5} X_i \to \mathbb{R}$ is a non-negative measurable function, then

$$\begin{aligned}
&\int_{\prod_{i=1}^{5} X_i} f \; d \bigotimes_{i=1}^{5} \lambda(x_1, \ldots, x_{i-1}).\mu_i(x_{par(i)}) \\
&= \int_{X_1} \left(\lambda x_1. \int_{X_2} \left(\lambda x_2. \int_{X_3} \left(\lambda x_3. \int_{X_4} \left(\lambda x_4. \int_{X_5} \lambda x_5. f(x_1, \ldots, x_5) \; d\mu_5(x_3) \right) \right. \right. \right. \\
&\qquad \left. \left. \left. d\mu_4(x_3) \right) \; d\mu_3(x_1, x_2) \right) \; d\mu_2 \right) \; d\mu_1.
\end{aligned}$$

Finally, it is common to need to evaluate integrals with respect to probability measures that are fusions or products. For this, Monte Carlo integration (Proposition A.5.3) is often appropriate, since an analytical solution may not be possible.

First consider products. Suppose that $\mu_1 : \mathcal{P}(X_1)$ is a probability measure, $\mu_2 : X_1 \to \mathcal{P}(X_2)$ a probability kernel, and $f : X_1 \times X_2 \to \mathbb{R}$ a non-negative, measurable function. The task is to (approximately) evaluate

$$\int_{X_1 \times X_2} f \; d(\mu_1 \otimes \mu_2).$$

Let $(\eta_n : \Omega \to X_1 \times X_2)_{n \in \mathbb{N}}$ be an i.i.d. sequence of random variables with distribution $\mu_1 \otimes \mu_2$. According to Proposition A.5.3, almost-surely,

$$\int_{X_1 \times X_2} f \; d(\mu_1 \otimes \mu_2) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} (f \circ \eta_j).$$

Now $\eta_n = (\eta_n^{(1)}, \eta_n^{(2)})$, where $\eta_n^{(1)} : \Omega \to X_1$ and $\eta_n^{(2)} : \Omega \to X_2$, for all $n \in \mathbb{N}$. Hence, almost surely,

$$\int_{X_1 \times X_2} f \; d(\mu_1 \otimes \mu_2) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} (f \circ (\eta_j^{(1)}, \eta_j^{(2)})),$$

where, by Proposition A.7.2, $\mathcal{L}(\eta_n^{(1)}) = \mu_1$ and $\mathcal{L}(\eta_n^{(2)}) = \mu_1 \odot \mu_2$, for all $n \in \mathbb{N}$. Informally,

$$\int_{X_1 \times X_2} f \; d(\mu_1 \otimes \mu_2) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} f(x_j^{(1)}, x_j^{(2)}),$$

where $x_n^{(1)} \sim \mu_1$ and $x_n^{(2)} \sim \mu_1 \odot \mu_2$, for all $n \in \mathbb{N}$.

Now consider fusions. Suppose that $\mu_1 : \mathcal{P}(X_1)$ is a probability measure, $\mu_2 : X_1 \to \mathcal{P}(X_2)$ a probability kernel, and $f : X_2 \to \mathbb{R}$ a non-negative, measurable function. The task is to (approximately) evaluate

$$\int_{X_2} f \; d(\mu_1 \odot \mu_2).$$

Let $(\xi_n : \Omega \to X_2)_{n\in\mathbb{N}}$ be an i.i.d. sequence of random variables with distribution $\mu_1 \odot \mu_2$. According to Proposition A.5.3, almost-surely,

$$\int_{X_2} f \; d(\mu_1 \odot \mu_2) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (f \circ \xi_j).$$

Informally,

$$\int_{X_1} f \; d(\mu_1 \odot \mu_2) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} f(x_j^{(2)}),$$

where $x_n^{(2)} \sim \mu_1 \odot \mu_2$, for all $n \in \mathbb{N}$.

The preceding discussion begs the question: how does one sample from fusion and product measures? Suppose that it is known how to sample from $\mu_1 : \mathcal{P}(X_1)$ and $\mu_2(x_1) : \mathcal{P}(X_2)$, for all $x_1 \in X_1$. Consider first the fusion probability measure $\mu_1 \odot \mu_2$. By definition,

$$\mu_1 \odot \mu_2 = \lambda A_2. \int_{X_1} \lambda x_1.\mu_2(x_1)(A_2) \; d\mu_1.$$

Then the algorithm for sampling from $\mu_1 \odot \mu_2$ is as follows:

sample $x_1 \sim \mu_1$;
sample $x_2 \sim \mu_2(x_1)$;
return x_2;

Now consider the product probability measure $\mu_1 \otimes \mu_2$. By definition,

$$(\mu_1 \otimes \mu_2) = \lambda A. \int_{X_1} \left(\lambda x_1. \int_{X_2} \lambda x_2.\mathbf{1}_A(x_1, x_2) \; d\mu_2(x_1) \right) \; d\mu_1.$$

Then the algorithm for sampling from $\mu_1 \otimes \mu_2$ is as follows:

sample $x_1 \sim \mu_1$;
sample $x_2 \sim \mu_2(x_1)$;
return (x_1, x_2);

To see that the two preceding algorithms are correct, consider the following. Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ measurable spaces, $\mu_1 : \mathcal{P}(X_1)$ a probability measure, $\mu_2 : X_1 \to \mathcal{P}(X_2)$ a probability kernel, and $f : \Omega \to X_1 \times X_2$ a random variable such that $\mathcal{L}(f) = \mu_1 \otimes \mu_2$. Suppose that $f = (f_1, f_2)$, where $f_1 : \Omega \to X_1$ and $f_2 : \Omega \to X_2$. Then, according to Proposition A.7.2, $\mathcal{L}(f_1) = \mu_1$ and $\mathcal{L}(f_2) = \mu_1 \odot \mu_2 = \lambda A_2.\mathsf{E}(\lambda\omega.\mu_2(f_1(\omega))(A_2))$, which justifies each algorithm.

A.8 Infinite Products of Probability Kernels

Now products of infinitely many probability kernels are considered. These can be used to model probability kernels whose codomain is probability measures over sequences, sets, or multisets. Here is a standard result about the existence of probability measures on an infinite product of spaces.

Notation. Let $\pi_{j_1,\ldots,j_k} : \prod_{n\in\mathbb{N}} X_n \to X_{j_1} \times \cdots \times X_{j_k}$ be the canonical projection, where $j_1, \ldots, j_k \in \mathbb{N}$ and $j_1 < \cdots < j_k$. If $J = \{j_1, \ldots, j_k\}$, then π_J means $\pi_{j_1,\ldots,j_k}$.

Proposition A.8.1. *(Ionescu-Tulcea theorem) Let $(X_n, \mathcal{A}_n)$ be a measurable space and $\mu_n : \prod_{j=1}^{n-1} X_j \to \mathcal{P}(X_n)$ a probability kernel, for all $n \in \mathbb{N}$. Then there exists a unique probability measure μ on $(\prod_{n\in\mathbb{N}} X_n, \bigotimes_{n\in\mathbb{N}} \mathcal{A}_n)$ such that $\mu \circ \pi_{1,\ldots,n}^{-1} = \bigotimes_{j=1}^{n} \mu_j$, for all $n \in \mathbb{N}$.*

Proof. See [87, Theorem 14.32], [83, Theorem 6.17], or [24, Theorem 4.7]. □

Notation. The unique probability measure μ given by Proposition A.8.1 is denoted by $\bigotimes_{n\in\mathbb{N}} \mu_n$.

By replacing each probability kernel μ_n in Proposition A.8.1 by a probability measure $\mu_n : \mathcal{P}(X_n)$, a corollary of Proposition A.8.1 is immediately obtained for a countably infinite product of probability measures. In this case, it is easy to see that $\mu \circ \pi_J^{-1} = \bigotimes_{j\in J} \mu_j$, for all (nonempty) finite subsets $J \subseteq \mathbb{N}$.

Proposition A.8.1 can be applied to function spaces, since a function space can be identified with a product for which each factor is the same. Thus consider a function space X^Y, where $Y \triangleq \{y_n\}_{n\in\mathbb{N}}$ is countably infinite. Let $X_y \triangleq X$, for all $y \in Y$. Then X^Y can be identified with $\prod_{n\in\mathbb{N}} X_{y_n}$, as measurable spaces. Suppose there exists a probability kernel $\mu_n : X^{\{y_1,\ldots,y_{n-1}\}} \to \mathcal{P}(X_{y_n})$, for all $n \in \mathbb{N}$. Then, by the result of Ionescu-Tulcea, there exists a unique probability measure μ on X^Y such that $\mu \circ \pi_{1,\ldots,n}^{-1} = \bigotimes_{j=1}^{n} \mu_j$, for all $n \in \mathbb{N}$.

Example A.8.1. Proposition A.8.1 can be used to construct a sequence of independent random variables of given distributions. For all $n \in \mathbb{N}$, let $(\Omega_n, \mathfrak{S}_n, \mathsf{P}_n)$ be a probability space, $(X_n, \mathcal{A}_n)$ a measurable space, and $f_n : \Omega_n \to X_n$ a random variable. Even when the $(\Omega_n, \mathfrak{S}_n, \mathsf{P}_n)$ are identical, the sequence $(f_n)_{n\in\mathbb{N}}$ may not be independent. Let $\Omega \triangleq \prod_{n\in\mathbb{N}} \Omega_n$ and $\mathfrak{S} \triangleq \bigotimes_{n\in\mathbb{N}} \mathfrak{S}_n$. By the corollary of Proposition A.8.1 discussed above, there exists a unique probability measure $\mathsf{P} \triangleq \bigotimes_{n\in\mathbb{N}} \mathsf{P}_n$ on $(\Omega, \mathfrak{S})$ such that $\mathsf{P} \circ \pi_J^{-1} = \bigotimes_{j\in J} \mathsf{P}_j$, for all finite subsets $J \subseteq \mathbb{N}$. This completes the construction of the probability space $(\Omega, \mathfrak{S}, \mathsf{P})$.

Now, for all $n \in \mathbb{N}$, define $\widetilde{f}_n : \Omega \to X_n$ by $\widetilde{f}_n = f_n \circ \pi_n$. Clearly, each $\widetilde{f}_n$ is measurable. Furthermore, $\mathsf{P} \circ \widetilde{f}_n^{-1} = \mathsf{P} \circ (f_n \circ \pi_n)^{-1} = (\mathsf{P} \circ \pi_n^{-1}) \circ f_n^{-1} = \mathsf{P}_n \circ f_n^{-1}$, for all $n \in \mathbb{N}$. Hence each $\widetilde{f}_n$ has the same distribution as f_n.

It remains to show that $(\widetilde{f}_n)_{n\in\mathbb{N}}$ is independent. Let J be a finite subset of $\mathbb{N}$ and $A_j \in \mathfrak{S}_j$, for all $j \in J$. Then

$$
\begin{aligned}
&\mathsf{P}\Big(\bigcap_{j\in J} \widetilde{f}_j^{-1}(A_j)\Big) \\
&= \mathsf{P}\Big(\bigcap_{j\in J} \pi_j^{-1}(f_j^{-1}(A_j))\Big) \\
&= \mathsf{P}\Big(\pi_J^{-1}\Big(\prod_{j\in J} f_j^{-1}(A_j)\Big)\Big) \\
&= \Big(\bigotimes_{j\in J} \mathsf{P}_j\Big)\Big(\prod_{j\in J} f_j^{-1}(A_j)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \prod_{j\in J} \mathsf{P}_j(f_j^{-1}(A_j)) \\
&= \prod_{j\in J} \mathsf{P}(\pi_j^{-1}(f_j^{-1}(A_j))) \\
&= \prod_{j\in J} \mathsf{P}(\widetilde{f}_j^{-1}(A_j)).
\end{aligned}
$$

Thus $(\widetilde{f}_n)_{n\in\mathbb{N}}$ is independent.

The next result shows that each μ_n in Proposition A.8.1 is a regular conditional distribution with respect to the probability measure $\bigotimes_{m\in\mathbb{N}} \mu_m$.

Proposition A.8.2. *Let $(X_n, \mathcal{A}_n)$ be a measurable space and $\mu_n : \prod_{j=1}^{n-1} X_j \to \mathcal{P}(X_n)$ a probability kernel, for all $n \in \mathbb{N}$. Let $\pi_{1,\ldots,n} : \prod_{m\in\mathbb{N}} X_m \to \prod_{j=1}^{n} X_j$ and $\pi_n : \prod_{m\in\mathbb{N}} X_m \to X_n$ be the canonical projections. Then, for all $n \in \mathbb{N}$ and $A_n \in \mathcal{A}_n$,*

$$
(\bigotimes_{m\in\mathbb{N}} \mu_m)(\pi_n^{-1}(A_n) \mid \pi_{1,\ldots,n-1}) = \lambda x.\mu_n(\pi_{1,\ldots,n-1}(x))(A_n) \quad \bigotimes_{m\in\mathbb{N}} \mu_m\text{-a.e.}
$$

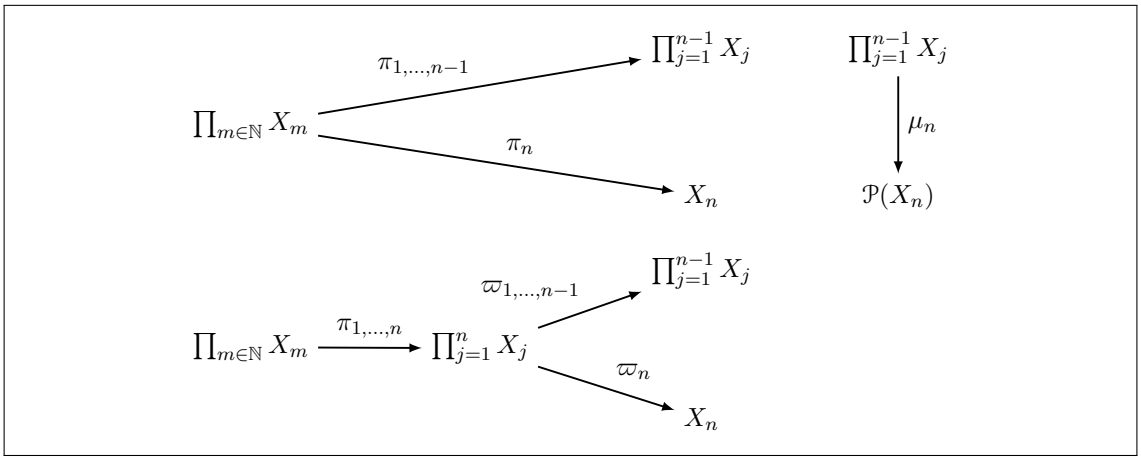

Figure A.30: Setting for Proposition A.8.2

Proof. Let $\varpi_{1,\ldots,n-1} : \prod_{j=1}^{n} X_j \to \prod_{j=1}^{n-1} X_j$ and $\varpi_n : \prod_{j=1}^{n} X_j \to X_n$ be the canonical projections. For all $n \in \mathbb{N}$ and $A_n \in \mathcal{A}_n$, $\bigotimes_{m\in\mathbb{N}} \mu_m$-almost surely,

$$
\begin{aligned}
&\lambda x.\mu_n(\pi_{1,\ldots,n-1}(x))(A_n) \\
&= \lambda y.\mu_n(\varpi_{1,\ldots,n-1}(y))(A_n) \circ \pi_{1,\ldots,n} \\
&= (\bigotimes_{j=1}^{n} \mu_j)(\varpi_n^{-1}(A_n) \mid \varpi_{1,\ldots,n-1}) \circ \pi_{1,\ldots,n} && \text{[Proposition A.7.17]} \\
&= ((\bigotimes_{m\in\mathbb{N}} \mu_m) \circ \pi_{1,\ldots,n}^{-1})(\varpi_n^{-1}(A_n) \mid \varpi_{1,\ldots,n-1}) \circ \pi_{1,\ldots,n} \\
&= (\bigotimes_{m\in\mathbb{N}} \mu_m)((\varpi_n \circ \pi_{1,\ldots,n})^{-1}(A_n) \mid \varpi_{1,\ldots,n-1} \circ \pi_{1,\ldots,n}) && \text{[Proposition A.5.13]}
\end{aligned}
$$

$$= (\bigotimes_{m\in\mathbb{N}} \mu_m)(\pi_n^{-1}(A_n) \mid \pi_{1,\dots,n-1}).$$

□

Now comes a useful generalization of Proposition A.8.1.

Notation. Extending previous notation, let $\pi_{j_1,\dots,j_k} : \prod_{j=m}^{\infty} X_j \to X_{j_1} \times \cdots \times X_{j_k}$ be the canonical projection, where $\{j_1, \dots, j_k\} \subseteq \{m, m+1, \dots\}$ and $j_1 < \cdots < j_k$. If $J = \{j_1, \dots, j_k\}$, then π_J means $\pi_{j_1,\dots,j_k}$.

Proposition A.8.3. *(Generalized Ionescu-Tulcea theorem) Let $(X_n, \mathcal{A}_n)$ be a measurable space and $\mu_n : \prod_{j=1}^{n-1} X_j \to \mathcal{P}(X_n)$ a probability kernel, for all $n \in \mathbb{N}$. Then, for all $m \in \mathbb{N}$, there exists a unique probability kernel $\bigotimes_{j=m}^{\infty} \mu_j : \prod_{j=1}^{m-1} X_j \to \mathcal{P}(\prod_{j=m}^{\infty} X_j)$ such that $\lambda t.((\bigotimes_{j=m}^{\infty} \mu_j)(t) \circ \pi_{m,\dots,n}^{-1}) = \bigotimes_{j=m}^{n} \mu_j$, for all $n \geq m$.*

Note. The Ionescu-Tulcea theorem can be recovered from Proposition A.8.3 by putting $m = 1$.

Proof. To do: Give this proof. □

The measure $\bigotimes_{j=m}^{\infty} \mu_j$ given by the generalized Ionescu-Tulcea theorem can be factored in a natural way.

Proposition A.8.4. *Let $(X_n, \mathcal{A}_n)$ be a measurable space and $\mu_n : \prod_{j=1}^{n-1} X_j \to \mathcal{P}(X_n)$ a probability kernel, for all $n \in \mathbb{N}$. Then, for all $m \in \mathbb{N}$ and $k \geq m$,*

$$\bigotimes_{j=m}^{\infty} \mu_j = (\bigotimes_{j=m}^{k} \mu_j) \otimes (\bigotimes_{j=k+1}^{\infty} \mu_j).$$

Note. Strictly speaking, the equality of the probability kernels stated in Proposition A.8.4 relies on the identification of the spaces $\prod_{j=m}^{\infty} X_j$ and $\prod_{j=m}^{k} X_j \times \prod_{j=k+1}^{\infty} X_j$.

Proof. By Proposition A.8.3, it suffices to show that, for all $m \in \mathbb{N}$ and $k, n \geq m$,

$$\lambda t.(((\bigotimes_{j=m}^{k} \mu_j) \otimes (\bigotimes_{j=k+1}^{\infty} \mu_j))(t) \circ \pi_{m,\dots,n}^{-1}) = \bigotimes_{j=m}^{n} \mu_j.$$

There are two cases to consider. Suppose first that $n \leq k$. Let $\varpi_{m,\dots,n} : \prod_{j=m}^{k} X_j \to \prod_{j=m}^{n} X_j$ denote the canonical projection, for all $m \in \mathbb{N}$ and $n \geq m$. Then, for all

$x_1 \in X_1, \ldots, x_{m-1} \in X_{m-1}$, and $B \in \bigotimes_{j=m}^{n} \mathcal{A}_j$,

$$
\begin{aligned}
&((\bigotimes_{j=m}^{k} \mu_j) \otimes (\bigotimes_{j=k+1}^{\infty} \mu_j))(x_1, \ldots, x_{m-1})(\pi_{m,\ldots,n}^{-1}(B)) \\
&= \int_{\prod_{j=m}^{k} X_j} \left(\lambda(x_m, \ldots, x_k). \int_{\prod_{j=k+1}^{\infty} X_j} \lambda(x_{k+1}, \ldots). \mathbf{1}_{\pi_{m,\ldots,n}^{-1}(B)}(x_m, \ldots) \right. \\
&\qquad \left. d(\bigotimes_{j=k+1}^{\infty} \mu_j)(x_1, \ldots, x_k) \right) \; d(\bigotimes_{j=m}^{k} \mu_j)(x_1, \ldots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{k} X_j} \left(\lambda(x_m, \ldots, x_k). \mathbf{1}_{\varpi_{m,\ldots,n}^{-1}(B)}(x_m, \ldots, x_k) \int_{\prod_{j=k+1}^{\infty} X_j} \right. \\
&\qquad \left. d(\bigotimes_{j=k+1}^{\infty} \mu_j)(x_1, \ldots, x_k) \right) \; d(\bigotimes_{j=m}^{k} \mu_j)(x_1, \ldots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{k} X_j} \lambda(x_m, \ldots, x_k). \mathbf{1}_{\varpi_{m,\ldots,n}^{-1}(B)}(x_m, \ldots, x_k) \; d(\bigotimes_{j=m}^{k} \mu_j)(x_1, \ldots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{n} X_j} \left(\lambda(x_m, \ldots, x_n). \int_{\prod_{j=n+1}^{k} X_j} \lambda(x_{n+1}, \ldots, x_k). \mathbf{1}_{\varpi_{m,\ldots,n}^{-1}(B)}(x_m, \ldots, x_k) \right. \\
&\qquad \left. d(\bigotimes_{j=n+1}^{k} \mu_j)(x_1, \ldots, x_n) \right) \; d(\bigotimes_{j=m}^{n} \mu_j)(x_1, \ldots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{n} X_j} \left(\lambda(x_m, \ldots, x_n). \mathbf{1}_B(x_m, \ldots, x_n) \int_{\prod_{j=n+1}^{k} X_j} d(\bigotimes_{j=n+1}^{k} \mu_j)(x_1, \ldots, x_n) \right) \\
&\qquad d(\bigotimes_{j=m}^{n} \mu_j)(x_1, \ldots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{n} X_j} \mathbf{1}_B \; d(\bigotimes_{j=m}^{n} \mu_j)(x_1, \ldots, x_{m-1}) \\
&= (\bigotimes_{j=m}^{n} \mu_j)(x_1, \ldots, x_{m-1})(B).
\end{aligned}
$$

Now consider the case when $n > k$. Then, for all $x_1 \in X_1, \ldots, x_{m-1} \in X_{m-1}$, and $B \in \bigotimes_{j=m}^{n} \mathcal{A}_j$,

$$
((\bigotimes_{j=m}^{k} \mu_j) \otimes (\bigotimes_{j=k+1}^{\infty} \mu_j))(x_1, \ldots, x_{m-1})(\pi_{m,\ldots,n}^{-1}(B))
$$

$$
\begin{aligned}
&= \int_{\prod_{j=m}^{k} X_j} \left(\lambda(x_m, \dots, x_k). \int_{\prod_{j=k+1}^{\infty} X_j} \lambda(x_{k+1}, \dots).\mathbf{1}_{\pi^{-1}_{m,\dots,n}(B)}(x_m, \dots) \right. \\
&\qquad \left. d(\bigotimes_{j=k+1}^{\infty} \mu_j)(x_1, \dots, x_k) \right) \; d(\bigotimes_{j=m}^{k} \mu_j)(x_1, \dots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{k} X_j} \left(\lambda(x_m, \dots, x_k). \int_{\prod_{j=k+1}^{\infty} X_j} \lambda(x_{k+1}, \dots, x_n).\mathbf{1}_B(x_m, \dots, x_n) \circ \pi_{k+1,\dots,n} \right. \\
&\qquad \left. d(\bigotimes_{j=k+1}^{\infty} \mu_j)(x_1, \dots, x_k) \right) \; d(\bigotimes_{j=m}^{k} \mu_j)(x_1, \dots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{k} X_j} \left(\lambda(x_m, \dots, x_k). \int_{\prod_{j=k+1}^{n} X_j} \lambda(x_{k+1}, \dots, x_n).\mathbf{1}_B(x_m, \dots, x_n) \right. \\
&\qquad \left. d(\bigotimes_{j=k+1}^{\infty} \mu_j)(x_1, \dots, x_k) \circ \pi^{-1}_{k+1,\dots,n} \right) \; d(\bigotimes_{j=m}^{k} \mu_j)(x_1, \dots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{k} X_j} \left(\lambda(x_m, \dots, x_k). \int_{\prod_{j=k+1}^{n} X_j} \lambda(x_{k+1}, \dots, x_n).\mathbf{1}_B(x_m, \dots, x_n) \right. \\
&\qquad \left. d(\bigotimes_{j=k+1}^{n} \mu_j)(x_1, \dots, x_k) \right) \; d(\bigotimes_{j=m}^{k} \mu_j)(x_1, \dots, x_{m-1}) \\
&= \int_{\prod_{j=m}^{n} X_j} \mathbf{1}_B \; d(\bigotimes_{j=m}^{n} \mu_j)(x_1, \dots, x_{m-1}) \\
&= (\bigotimes_{j=m}^{n} \mu_j)(x_1, \dots, x_{m-1})(B).
\end{aligned}
$$

□

Here is the analogue of Proposition A.7.21 for infinite products.

Proposition A.8.5. *Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$ and $(Z_n, \mathcal{C}_n)$, for all $n \in \mathbb{N}$, be measurable spaces and $\mu_n : X \times Y \times \prod_{j=1}^{n-1} Z_j \to \mathcal{P}(Z_n)$, for all $n \in \mathbb{N}$, probability kernels. Then, for all $x \in X$,*

$$\lambda y.(\bigotimes_{n\in\mathbb{N}} \mu_n)(x, y) = \bigotimes_{n\in\mathbb{N}} \lambda(y, z_1, \dots, z_{n-1}).\mu_n(x, y, z_1, \dots, z_{n-1}).$$

Proof. By Proposition A.8.3, $\bigotimes_{n\in\mathbb{N}} \mu_n : X \times Y \to \mathcal{P}(\prod_{n\in\mathbb{N}} Z_n)$ is the unique probability kernel such that $\lambda(x, y).((\bigotimes_{n\in\mathbb{N}} \mu_n)(x, y) \circ \pi^{-1}_{1,\dots,m}) = \bigotimes_{j=1}^{m} \mu_j$, for all $m \in \mathbb{N}$.

Also, by Proposition A.8.3, for all $x \in X$, $\bigotimes_{n\in\mathbb{N}} \lambda(y, z_1, \dots, z_{n-1}).\mu_n(x, y, z_1, \dots, z_{n-1}) :$

$Y \to \mathcal{P}(\prod_{n\in\mathbb{N}} Z_n)$ is the unique probability kernel such that

$$\lambda t.((\bigotimes_{n\in\mathbb{N}} \lambda(y, z_1, \ldots, z_{n-1}).\mu_i(x, y, z_1, \ldots, z_{n-1}))(t) \circ \pi^{-1}_{1,\ldots,m}) = \bigotimes_{j=1}^{m} \lambda(y, z_1, \ldots, z_{j-1}).\mu_j(x, y, z_1, \ldots, z_{j-1}),$$

for all $m \in \mathbb{N}$.

Now, for all $x \in X$ and $m \in \mathbb{N}$,

$$\begin{aligned}
&\bigotimes_{j=1}^{m} \lambda(y, z_1, \ldots, z_{j-1}).\mu_j(x, y, z_1, \ldots, z_{j-1}) \\
&= \lambda y.(\bigotimes_{j=1}^{m} \mu_j)(x, y) && \text{[Proposition A.7.21]} \\
&= \lambda y.(\lambda(x, y).((\bigotimes_{n\in\mathbb{N}} \mu_n)(x, y) \circ \pi^{-1}_{1,\ldots,m}))(x, y) && \text{[Proposition A.8.3]} \\
&= \lambda y.((\bigotimes_{n\in\mathbb{N}} \mu_n)(x, y) \circ \pi^{-1}_{1,\ldots,m}) \\
&= \lambda t.((\lambda y.(\bigotimes_{n\in\mathbb{N}} \mu_n)(x, y))(t) \circ \pi^{-1}_{1,\ldots,m}).
\end{aligned}$$

Thus, for all $x \in X$, $\lambda y.(\bigotimes_{n\in\mathbb{N}} \mu_n)(x, y) = \bigotimes_{n\in\mathbb{N}} \lambda(y, z_1, \ldots, z_{n-1}).\mu_n(x, y, z_1, \ldots, z_{n-1})$, by Proposition A.8.3. □

The next result states that if μ_n is a regular conditional distribution of f_n given $(f_0, \ldots, f_{n-1})$, for all $n \in \mathbb{N}$, then $\bigotimes_{n\in\mathbb{N}} \mu_n$ is a regular conditional distribution of $(f_1, f_2, \ldots)$ given f_0. This is the analogue for infinite products of Proposition A.7.12.

Notation. If $f : Y \to \prod_{n\in\mathbb{N}} X_n$, then $(f_1, \ldots, f_n) : Y \to \prod_{j=1}^{n} X_j$ is defined by $(f_1, \ldots, f_n) = \pi_{1,\ldots,n} \circ f$, for all $n \in \mathbb{N}$.

Proposition A.8.6. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space and $(X_n, \mathcal{A}_n)$ a measurable space, for all $n \in \mathbb{N}_0$. Suppose that $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{n\in\mathbb{N}} X_n$ are measurable, and, for all $n \in \mathbb{N}$, $\mu_n : \prod_{j=0}^{n-1} X_j \to \mathcal{P}(X_n)$ is a probability kernel such that*

$$\mathsf{P}(f_n^{-1}(A_n) \mid (f_0, \ldots, f_{n-1})) = \lambda\omega.\mu_n((f_0, \ldots, f_{n-1})(\omega))(A_n) \text{ a.s.},$$

for all $A_n \in \mathcal{A}_n$. Then

$$\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{n\in\mathbb{N}} \mu_n)(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{n\in\mathbb{N}} \mathcal{A}_n$.

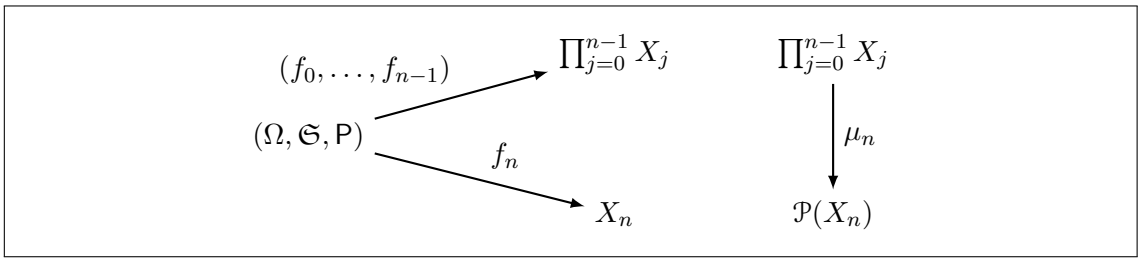

Figure A.31: Setting for Proposition A.8.6

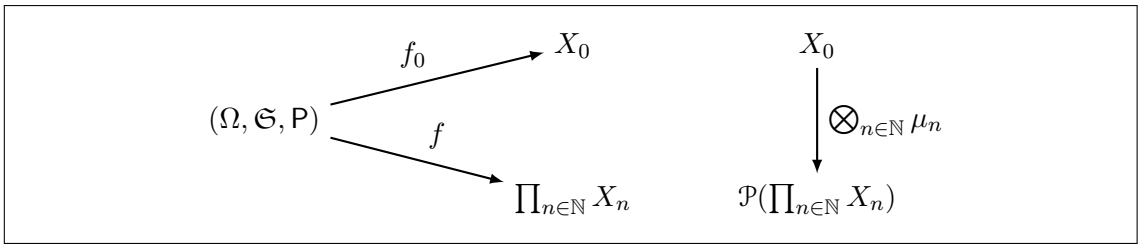

Figure A.32: Setting for Proposition A.8.6

Proof. Note that there exist measurable functions $f_n : \Omega \to X_n$, for all $n \in \mathbb{N}$ such that $f = (f_1, f_2, \ldots)$. By Proposition A.8.3, $\bigotimes_{n\in\mathbb{N}} \mu_n : X_0 \to \mathcal{P}(\prod_{n\in\mathbb{N}} X_n)$ exists and satisfies $\lambda t.((\bigotimes_{n\in\mathbb{N}} \mu_n)(t) \circ \pi^{-1}_{1,\ldots,m}) = \bigotimes_{j=1}^{m} \mu_j$, for all $m \in \mathbb{N}$.

Let

$$\mathcal{P} \triangleq \{\prod_{i=1}^{m} A_i \times \prod_{i=m+1}^{\infty} X_i \mid A_i \in \mathcal{A}_i, \text{ for } i = 1, \ldots, m \text{ and for all } m \in \mathbb{N}\}$$

and

$$\mathcal{L} \triangleq \{A \in \bigotimes_{n\in\mathbb{N}} \mathcal{A}_n \mid \mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{n\in\mathbb{N}} \mu_n)(f_0(\omega))(A) \text{ a.s.}\}.$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \bigotimes_{n\in\mathbb{N}} \mathcal{A}_n$.

Suppose that $A \triangleq \prod_{i=1}^{m} A_i \times \prod_{i=m+1}^{\infty} X_i \in \mathcal{P}$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(A) \mid f_0) \\
&= \mathsf{P}((f_1, \ldots, f_m)^{-1}(A_1 \times \cdots \times A_m) \mid f_0) \\
&= \lambda\omega.(\bigotimes_{j=1}^{m} \mu_j)(f_0(\omega))(A_1 \times \cdots \times A_m) && \text{[Proposition A.7.12]} \\
&= \lambda\omega.((\bigotimes_{n\in\mathbb{N}} \mu_n)(f_0(\omega)) \circ \pi^{-1}_{1,\ldots,m})(A_1 \times \cdots \times A_m) && \text{[Proposition A.8.3]} \\
&= \lambda\omega.(\bigotimes_{n\in\mathbb{N}} \mu_n)(f_0(\omega))(A).
\end{aligned}$$

Thus $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, clearly $\prod_{n\in\mathbb{N}} X_n \in \mathcal{L}$. Second, suppose that $(A_k)_{k\in\mathbb{N}}$ is a increasing sequence in $\mathcal{L}$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(\bigcup_{k\in\mathbb{N}} A_k) \mid f_0) \\
&= \mathsf{P}(\bigcup_{k\in\mathbb{N}} f^{-1}(A_k) \mid f_0) \\
&= \mathsf{E}(\lim_{k\to\infty} \mathbf{1}_{f^{-1}(A_k)} \mid f_0) \\
&= \lim_{k\to\infty} \mathsf{E}(\mathbf{1}_{f^{-1}(A_k)} \mid f_0) && \text{[Proposition A.5.8]} \\
&= \lim_{k\to\infty} \lambda\omega.(\bigotimes_{n\in\mathbb{N}} \mu_i)(f_0(\omega))(A_k) \\
&= \lambda\omega.(\bigotimes_{n\in\mathbb{N}} \mu_i)(f_0(\omega))(\bigcup_{k\in\mathbb{N}} A_k).
\end{aligned}$$

Thus $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{L}$. Third, suppose that $A, B \in \mathcal{L}$ and $A \subseteq B$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(B \setminus A) \mid f_0) \\
&= \mathsf{P}(f^{-1}(B) \setminus f^{-1}(A) \mid f_0) \\
&= \mathsf{E}(\mathbf{1}_{f^{-1}(B)} - \mathbf{1}_{f^{-1}(A)} \mid f_0) && [A \subseteq B] \\
&= \mathsf{P}(f^{-1}(B) \mid f_0) - \mathsf{P}(f^{-1}(A) \mid f_0) && \text{[Proposition A.5.6, Part 1]} \\
&= \lambda\omega.(\bigotimes_{n\in\mathbb{N}} \mu_i)(f_0(\omega))(B) - \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A) \\
&= \lambda\omega.(\bigotimes_{n\in\mathbb{N}} \mu_i)(f_0(\omega))(B \setminus A).
\end{aligned}$$

Thus $B \setminus A \in \mathcal{L}$. Hence it has been shown that $\mathcal{L}$ is a λ-system.

By Proposition A.1.2, $\sigma(\mathcal{P}) \subseteq \mathcal{L}$; that is, $\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{n\in\mathbb{N}} \mu_i)(f_0(\omega))(A)$ a.s., for all $A \in \bigotimes_{i=1}^{n} \mathcal{A}_i$. □

The next result goes in the converse direction to Proposition A.8.1.

Proposition A.8.7. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_n)_{n\in\mathbb{N}}$ a sequence of standard Borel spaces, and $\Gamma : \Omega \to \prod_{n\in\mathbb{N}} X_n$ a stochastic process. Then there exists a sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability kernels such that $\mu_n : \prod_{j=1}^{n-1} X_j \to \mathcal{P}(X_n)$ and $\mathcal{L}((\Gamma_1, \ldots, \Gamma_n)) = \bigotimes_{j=1}^{n} \mu_j$, for all $n \in \mathbb{N}$.*

Proof. The sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability kernels is defined by induction. Let $\mu_1 \triangleq \mathcal{L}(\Gamma_1)$. Then $\mu_1 \in \mathcal{P}(X_1)$, as required.

Now suppose that $\mu_1, \ldots, \mu_n$ are defined. Thus $\mathcal{L}((\Gamma_1, \ldots, \Gamma_n)) = \bigotimes_{j=1}^{n} \mu_j$. Note that $\mathcal{L}((\Gamma_1, \ldots, \Gamma_n)) = \mathcal{L}(\pi_{1,\ldots,n} \circ \Gamma) = \mathcal{L}(\Gamma) \circ \pi_{1,\ldots,n}^{-1} = \mathsf{P} \circ (\pi_{1,\ldots,n} \circ \Gamma)^{-1}$. Let $\pi_{n+1} : \prod_{n\in\mathbb{N}} X_n \to X_{n+1}$ be the canonical projection. By Proposition A.5.16, there exists a probability kernel $\mu_{n+1} : \prod_{j=1}^{n} X_j \to \mathcal{P}(X_{n+1})$ such that

$$\mathsf{P}((\pi_{n+1} \circ \Gamma)^{-1}(A_{n+1}) \,|\, \pi_{1,\ldots,n} \circ \Gamma) = \lambda\omega.\mu_{n+1}(\pi_{1,\ldots,n}(\Gamma(\omega)))(A_{n+1}) \text{ a.s.},$$

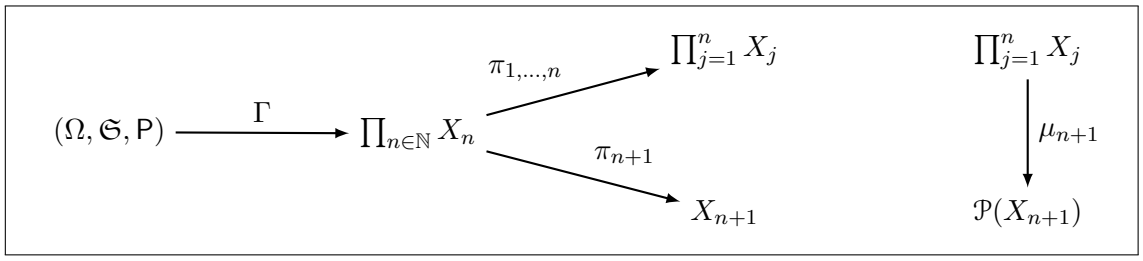

Figure A.33: Setting for Proposition A.8.7

for all $A_{n+1} \in \mathcal{A}_{n+1}$.

Then, for all $A \in \bigotimes_{j=1}^n \mathcal{A}_j$ and $A_{n+1} \in \mathcal{A}_{n+1}$,

$$
\begin{aligned}
&(\bigotimes_{j=1}^{n+1} \mu_j)(A \times A_{n+1}) \\
&= (\bigotimes_{j=1}^{n} \mu_j \otimes \mu_{n+1})(A \times A_{n+1}) \\
&= \int_{\prod_{j=1}^n X_j} \mathbf{1}_A \; \lambda x.\mu_{n+1}(x)(A_{n+1}) \; d \bigotimes_{j=1}^{n} \mu_j \\
&= \int_{\prod_{j=1}^n X_j} \mathbf{1}_A \; \lambda x.\mu_{n+1}(x)(A_{n+1}) \; d(\mathsf{P} \circ (\pi_{1,\dots,n} \circ \Gamma)^{-1}) && \text{[Induction hypothesis]} \\
&= \int_{\Omega} (\mathbf{1}_A \; \lambda x.\mu_{n+1}(x)(A_{n+1})) \circ (\pi_{1,\dots,n} \circ \Gamma) \; d\mathsf{P} && \text{[Proposition A.2.14]} \\
&= \int_{\Omega} \mathbf{1}_{(\pi_{1,\dots,n} \circ \Gamma)^{-1}(A)} \; \lambda\omega.\mu_{n+1}(\pi_{1,\dots,n}(\Gamma(\omega)))(A_{n+1}) \; d\mathsf{P} \\
&= \int_{\Omega} \mathbf{1}_{(\pi_{1,\dots,n} \circ \Gamma)^{-1}(A)} \; \mathsf{P}((\pi_{n+1} \circ \Gamma)^{-1}(A_{n+1}) | \pi_{1,\dots,n} \circ \Gamma) \; d\mathsf{P} \\
&= \int_{\Omega} \mathbf{1}_{(\pi_{1,\dots,n} \circ \Gamma)^{-1}(A)} \; \mathbf{1}_{(\pi_{n+1} \circ \Gamma)^{-1}(A_{n+1})} \; d\mathsf{P} \\
&= \int_{\Omega} \mathbf{1}_{(\pi_{1,\dots,n+1} \circ \Gamma)^{-1}(A \times A_{n+1})} \; d\mathsf{P} \\
&= (\mathsf{P} \circ (\pi_{1,\dots,n+1} \circ \Gamma)^{-1})(A \times A_{n+1}).
\end{aligned}
$$

Thus, by Proposition A.2.10, it follows that $\mathcal{L}((\Gamma_1, \dots, \Gamma_{n+1})) = \bigotimes_{j=1}^{n+1} \mu_j$. □

Now Proposition A.7.19 is extended to the case of an infinite product.

Proposition A.8.8. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, and $(X_n, \mathcal{A}_n)$ a standard Borel space, for all $n \in \mathbb{N}$. Suppose that $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{n\in\mathbb{N}} X_n$ are measurable. Let $\mu : X_0 \to \mathcal{P}(\prod_{n\in\mathbb{N}} X_n)$ be a regular conditional distribution of f given f_0. Then there exists a sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability kernels such that, for all $n \in \mathbb{N}$, $\mu_n : \prod_{i=0}^{n-1} X_i \to \mathcal{P}(X_n)$ is a regular conditional distribution of f_n given $(f_0, \dots, f_{n-1})$ and $\mu = \bigotimes_{n\in\mathbb{N}} \mu_n$ $\mathcal{L}(f_0)$-a.e.*

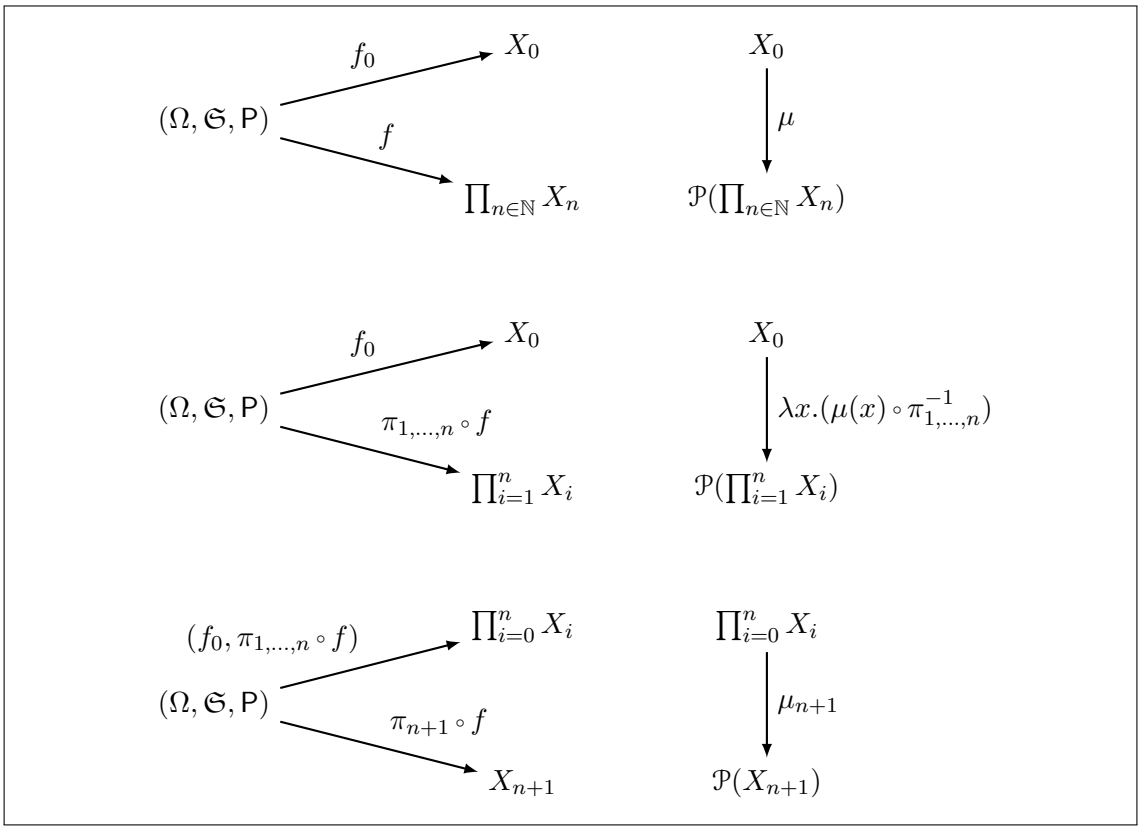

Figure A.34: Setting for Proposition A.8.8

Proof. Recall that $\pi_{1,\ldots,n} : \prod_{n\in\mathbb{N}} X_n \to \prod_{i=1}^n X_i$ is the canonical projection.

For all $n \in \mathbb{N}$, consider the probability kernel $\lambda x.(\mu(x) \circ \pi^{-1}_{1,\ldots,n}) : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$. It is shown that this probability kernel is a regular conditional distribution of $\pi_{1,\ldots,n} \circ f$ given f_0. Now, for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$, almost surely,

$$\begin{aligned}
&\mathsf{P}((\pi_{1,\ldots,n} \circ f)^{-1}(A) \mid f_0) \\
&= \mathsf{P}(f^{-1}(\pi^{-1}_{1,\ldots,n}(A)) \mid f_0) \\
&= \lambda\omega.\mu(f_0(\omega))(\pi^{-1}_{1,\ldots,n}(A)) && [\mu \text{ is a regular conditional distribution}] \\
&= \lambda\omega.(\mu(f_0(\omega)) \circ \pi^{-1}_{1,\ldots,n})(A) \\
&= \lambda\omega.\lambda x.(\mu(x) \circ \pi^{-1}_{1,\ldots,n})(f_0(\omega))(A).
\end{aligned}$$

Hence $\lambda x.(\mu(x) \circ \pi^{-1}_{1,\ldots,n})$ is a regular conditional distribution of $\pi_{1,\ldots,n} \circ f$ given f_0.

Now it is proved by induction that there exists a sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability kernels such that, for all $n \in \mathbb{N}$, $\mu_n : \prod_{i=0}^{n-1} X_i \to \mathcal{P}(X_n)$ is a regular conditional distribution of f_n given $(f_0, \ldots, f_{n-1})$ and $\lambda x.(\mu(x) \circ \pi^{-1}_{1,\ldots,n}) = \bigotimes_{i=1}^n \mu_i$ $\mathcal{L}(f_0)$-a.e.

Define $\mu_1 : X_0 \to \mathcal{P}(X_1)$ by $\mu_1 = \lambda x.(\mu(x) \circ \pi_1^{-1})$, which gives the result for $n = 1$.

Now assume the result holds for n. Thus there exists $\mu_i : \prod_{j=0}^{i-1} X_j \to \mathcal{P}(X_i)$ that is a regular conditional distribution of f_i given $(f_0, \ldots, f_{i-1})$, for $i = 1, \ldots, n$, such that

$\lambda x.(\mu(x) \circ \pi^{-1}_{1,\ldots,n}) = \bigotimes_{i=1}^{n} \mu_i \;\; \mathcal{L}(f_0)$-a.e. Consequently, for all $A \in \bigotimes_{i=1}^{n} \mathcal{A}_i$,

$$\mathsf{P}(\pi_{1,\ldots,n} \circ f)^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n} \mu_i)(f_0(\omega))(A) \text{ a.s.}$$

By Proposition A.5.16, there exists a probability kernel $\mu_{n+1} : \prod_{i=0}^{n} X_i \to \mathcal{P}(X_{n+1})$ such that, for all $A_{n+1} \in \mathcal{A}_{n+1}$,

$$\mathsf{P}((\pi_{n+1} \circ f)^{-1}(A_{n+1}) \mid (f_0, \pi_{1,\ldots,n} \circ f)) = \lambda\omega.\mu_{n+1}((f_0, \pi_{1,\ldots,n} \circ f)(\omega))(A_{n+1}) \text{ a.s.}$$

That is, μ_{n+1} is a regular conditional distribution of f_{n+1} given $(f_0, \ldots, f_n)$. Thus, by Proposition A.7.12, for all $B \in \bigotimes_{i=1}^{n+1} \mathcal{A}_i$,

$$\mathsf{P}((\pi_{1,\ldots,n+1} \circ f)^{-1}(B) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^{n+1} \mu_n)(f_0(\omega))(B) \text{ a.s.}$$

By the uniqueness part of Proposition A.5.16, it follows that

$$\lambda x.(\mu(x) \circ \pi^{-1}_{1,\ldots,n+1}) = \bigotimes_{i=1}^{n+1} \mu_i \;\; \mathcal{L}(f_0)\text{-a.e.}$$

This completes the induction argument.

It now follows from Proposition A.8.3 that $\mu = \bigotimes_{n\in\mathbb{N}} \mu_n \;\; \mathcal{L}(f_0)$-a.e. □

An important application of Proposition A.8.1 is to construct probability measures on the set of subsets of a set and also on the set of multisets on a set. First, sets are discussed. Recall that $\mathbb{B}^Y$ can be regarded variously as the set of subsets of Y or as the set of bit strings whose elements are indexed by elements of Y or as the product space consisting of $|Y|$ copies of $\mathbb{B}$. As usual, $\mathbb{B}^Y$ is given the σ-algebra generated by the evaluation maps.

In the following discussion, Y is assumed to be countably infinite. (If Y is finite, the changes needed are obvious.) Thus Y is isomorphic to $\mathbb{N}$ and has the form $(y_n)_{n\in\mathbb{N}}$.

Now let $F \triangleq \{y_1, \ldots, y_m\}$ be a (non-empty) finite subset of Y. For $k = 1, \ldots, m$, let

$$\mu_{y_k} : \mathbb{B}^{k-1} \to \mathcal{P}(\mathbb{B})$$

be a probability kernel. For $k > m$, define the probability kernel

$$\mu_{y_k} : \mathbb{B}^{k-1} \to \mathcal{P}(\mathbb{B})$$

by

$$\mu_{y_k}(b_1, \ldots, b_{k-1}) = \delta_{\mathsf{F}},$$

for all $(b_1, \ldots, b_{k-1}) \in \mathbb{B}^{k-1}$. (Here δ_{F} is the Dirac measure at F.) Now define μ to be the unique probability measure on $\mathbb{B}^Y$ such that $\mu \circ \pi^{-1}_{1,\ldots,n} = \bigotimes_{j=1}^{n} \mu_{y_j}$, for all $n \in \mathbb{N}$, guaranteed by Proposition A.8.1.

A crucial property of the probability measure μ thus constructed on $\mathbb{B}^Y$ is that an integral of the form $\int_{\mathbb{B}^Y} f \; d\mu$ can be reduced to an integral over the space $\mathbb{B}^F$. The technical details of this are given by the following proposition.

Proposition A.8.9. *Let $(Y, \mathcal{B})$ be a measurable space, where Y is countably infinite and enumerated as $(y_n)_{n\in\mathbb{N}}$, and $F \triangleq \{y_1, \ldots, y_m\}$ a finite subset of Y. For $k = 1, \ldots, m$, let*

$$\mu_{y_k} : \mathbb{B}^{k-1} \to \mathcal{P}(\mathbb{B})$$

be a probability kernel. For all $k > m$, define the probability kernel

$$\mu_{y_k} : \mathbb{B}^{k-1} \to \mathcal{P}(\mathbb{B})$$

by

$$\mu_{y_k}(b_1, \ldots, b_{k-1}) = \delta_{\mathsf{F}},$$

for all $(b_1, \ldots, b_{k-1}) \in \mathbb{B}^{k-1}$. Suppose that μ is the unique probability measure on $\mathbb{B}^Y$ such that $\mu \circ \pi^{-1}_{1,\ldots,n} = \bigotimes_{j=1}^n \mu_{y_j}$, for all $n \in \mathbb{N}$. Then the following hold.

1. *$\mu(\{s \in \mathbb{B}^Y \mid s(y) = \mathsf{F}, \text{ for all } y \in Y \setminus F\}) = 1$.*
2. *Let $f : \mathbb{B}^Y \to \mathbb{R}$ be integrable. Define $f' : \mathbb{B}^m \to \mathbb{R}$ by*

   $$f'(b_1, \ldots, b_m) = f(b_1, \ldots, b_m, \mathsf{F}, \mathsf{F}, \mathsf{F}, \ldots),$$

   for all $(b_1, \ldots, b_m) \in \mathbb{B}^m$. Then

   $$\int_{\mathbb{B}^Y} f \, d\mu = \int_{\mathbb{B}^m} f' \, d \bigotimes_{j=1}^m \mu_{y_j}.$$

Proof. 1. For all $n \in \mathbb{N}$, define $A_n \subseteq \mathbb{B}^n$ by

$$A_n = \{s \in \mathbb{B}^n \mid s(y) = \mathsf{F}, \text{ for all } y \in \{y_1, \ldots, y_n\} \setminus F\}.$$

Then, for $n \geq m$,

$$\begin{aligned}
&(\bigotimes_{i=1}^n \mu_{y_i})(A_n) \\
&= \int_{\mathbb{B}^n} \mathbf{1}_{A_n} \, d \bigotimes_{i=1}^n \mu_{y_i} \\
&= \int_{\mathbb{B}^n} \mathbf{1}_{\mathbb{B}^m} \, \mathbf{1}_{\{\mathsf{F}\}^{n-m}} \, d \bigotimes_{i=1}^n \mu_{y_i} \\
&= \int_{\mathbb{B}_m} \Bigg(\lambda(b_1, \ldots, b_m). \int_{\mathbb{B}^{n-m}} \lambda(b_{m+1}, \ldots, b_n). \mathbf{1}_{\mathbb{B}^m}(b_1, \ldots, b_m) \\
&\qquad\qquad \mathbf{1}_{\{\mathsf{F}\}^{n-m}}(b_{m+1}, \ldots b_n) \, d(\bigotimes_{i=m+1}^n \mu_{y_i})(b_1, \ldots, b_m) \Bigg) \, d \bigotimes_{i=1}^m \mu_{y_i} \\
&= \int_{\mathbb{B}^m} \Bigg(\lambda(b_1, \ldots, b_m). \mathbf{1}_{\mathbb{B}^m}(b_1, \ldots, b_m) \int_{\mathbb{B}^{n-m}} \lambda(b_{m+1}, \ldots, b_n). \\
&\qquad\qquad \mathbf{1}_{\{\mathsf{F}\}^{n-m}}(b_{m+1}, \ldots b_n) \, d(\bigotimes_{i=m+1}^n \mu_{y_i})(b_1, \ldots, b_m) \Bigg) \, d \bigotimes_{i=1}^m \mu_{y_i} \\
&= \int_{\mathbb{B}^m} \lambda(b_1, \ldots, b_m). \mathbf{1}_{\mathbb{B}^m}(b_1, \ldots, b_m) \, d \bigotimes_{i=1}^m \mu_{y_i} \\
&= 1.
\end{aligned}$$

That is, $(\bigotimes_{i=1}^{n} \mu_{y_i})(A_n) = 1$, for $n \geq m$.

Since $\mu \circ \pi_{1,\ldots,n}^{-1} = \bigotimes_{j=1}^{n} \mu_{y_j}$, for all $n \in \mathbb{N}$, it follows that $\mu(\pi_{1,\ldots,n}^{-1}(A_n)) = 1$, for $n \geq m$. But

$$\bigcap_{n \geq m} \pi_{1,\ldots,n}^{-1}(A_n) = \{s \in \mathbb{B}^Y \mid s(y) = \mathsf{F}, \text{ for all } y \in Y \setminus F\}.$$

Hence $\mu(\{s \in \mathbb{B}^Y \mid s(y) = \mathsf{F}, \text{ for all } y \in Y \setminus F\}) = 1$.

2. Letting $\mathcal{F} \triangleq \{s \in \mathbb{B}^Y \mid s(y) = \mathsf{F}, \text{ for all } y \in Y \setminus F\}$, it follows that

$$\begin{aligned}
&\int_{\mathbb{B}^Y} f \, d\mu \\
&= \int_{\mathbb{B}^Y} \mathbf{1}_{\mathcal{F}} \, f \, d\mu + \int_{\mathbb{B}^Y} \mathbf{1}_{\mathbb{B}^Y \setminus \mathcal{F}} \, f \, d\mu \\
&= \int_{\mathbb{B}^Y} \mathbf{1}_{\mathcal{F}} \, f \, d\mu \\
&= \int_{\mathbb{B}^Y} f' \circ \pi_{1,\ldots,m} \, d\mu \\
&= \int_{\mathbb{B}^m} f' \, d(\mu \circ \pi_{1,\ldots,m}^{-1}) \\
&= \int_{\mathbb{B}^m} f' \, d \bigotimes_{j=1}^{m} \mu_{y_j}.
\end{aligned}$$

□

In effect, f' is f restricted to the subsets of F.

As a special case of Proposition A.8.9, for measurable $A \subseteq \mathbb{B}^Y$, let $f \triangleq \mathbf{1}_A$. Define $A' \subseteq \mathbb{B}^m$ by $(b_1, \ldots, b_m) \in A'$ if and only if $(b_1, \ldots, b_m, \mathsf{F}, \mathsf{F}, \mathsf{F}, \ldots) \in A$. Then, for all measurable $A \subseteq \mathbb{B}^Y$, $\mu(A) = (\bigotimes_{i=1}^{m} \mu_i)(A')$.

The analogue for multisets of Proposition A.8.9 will be needed.

Proposition A.8.10. *Let $(Y, \mathcal{B})$ be a measurable space, where Y is countably infinite and enumerated as $(y_n)_{n \in \mathbb{N}}$, and $F \triangleq \{y_1, \ldots, y_m\}$ a finite subset of Y. For $k = 1, \ldots, m$, let*

$$\mu_{y_k} : \mathbb{N}_0^{k-1} \to \mathcal{P}(\mathbb{N}_0)$$

be a probability kernel. For all $k > m$, define the probability kernel

$$\mu_{y_k} : \mathbb{N}_0^{k-1} \to \mathcal{P}(\mathbb{N}_0)$$

by

$$\mu_{y_k}(n_1, \ldots, n_{k-1}) = \delta_0,$$

for all $(n_1, \ldots, n_{k-1}) \in \mathbb{N}_0^{k-1}$. Suppose that μ is the unique probability measure on $\mathbb{N}_0^Y$ such that $\mu \circ \pi_{1,\ldots,n}^{-1} = \bigotimes_{j=1}^{n} \mu_{y_j}$, for all $n \in \mathbb{N}$. Then the following hold.

1. *$\mu(\{s \in \mathbb{N}_0^Y \mid s(y) = 0, \text{ for all } y \in Y \setminus F\}) = 1$.*

2. *Let* $f : \mathbb{N}_0^Y \to \mathbb{R}$ *be integrable. Define* $f' : \mathbb{N}_0^m \to \mathbb{R}$ *by*

$$f'(n_1, \ldots, n_m) = f(n_1, \ldots, n_m, 0, 0, 0, \ldots),$$

for all $(n_1, \ldots, n_m) \in \mathbb{N}_0^m$. *Then*

$$\int_{\mathbb{N}_0^Y} f \, d\mu = \int_{\mathbb{N}_0^m} f' \, d \bigotimes_{j=1}^{m} \mu_{y_j}.$$

Proof. The proof is analogous to the proof of Proposition A.8.9. □

A.9 Sums of Probability Kernels

For sums, it is possible to do the finite and countably infinite cases together. Throughout this section, the index set I is assumed to be countable.

Given probability kernels $\mu_i : X \to \mathcal{P}(X_i)$, for all $i \in I$, the first task is to define a probability kernel $\bigoplus_{i \in I} \mu_i : X \to \mathcal{P}(\coprod_{i \in I} X_i)$ that is the sum of $(\mu_i)_{i \in I}$. Perhaps unexpectedly, this turns out to be rather subtle. Instead of starting with $\mu_i : X \to \mathcal{P}(X_i)$, it is necessary to start instead with $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$, for all $i \in I$. (Recall that $X_i \sqcup \{*\}$ is the one point extension of X_i; see Example A.1.1.) It will become apparent from Propositions A.9.2, A.9.3, and A.9.4 below why this is necessary. In addition, a deeper understanding of this will be given in Section A.10 after the concept of a quotient probability kernel has been introduced.

Definition A.9.1. Let $(X, \mathcal{A})$ be a measurable space and, for all $i \in I$, $(X_i, \mathcal{A}_i)$ a measurable space and $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ a probability kernel. Suppose that $\sum_{i \in I} \mu_i(x)(X_i) = 1$, for all $x \in X$. Then the *sum* $\bigoplus_{i \in I} \mu_i : X \to \mathcal{P}(\coprod_{i \in I} X_i)$ of $(\mu_i)_{i \in I}$ is defined by

$$(\bigoplus_{i \in I} \mu_i)(x)(\coprod_{i \in I} A_i) = \sum_{i \in I} \mu_i(x)(A_i),$$

for all $x \in X$ and $\coprod_{i \in I} A_i \in \bigoplus_{i \in I} \mathcal{A}_i$. Each μ_i is called an *addendum* of the sum.

Proposition A.9.1. *Let* $(X, \mathcal{A})$ *be a measurable space and, for all* $i \in I$, $(X_i, \mathcal{A}_i)$ *a measurable space and* $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ *a probability kernel. Suppose that* $\sum_{i \in I} \mu_i(x)(X_i) = 1$, *for all* $x \in X$. *Then* $\bigoplus_{i \in I} \mu_i : X \to \mathcal{P}(\coprod_{i \in I} X_i)$ *is a probability kernel.*

Proof. First, note that, for all $x \in X$, $(\bigoplus_{i \in I} \mu_i)(x)(\coprod_{i \in I} X_i) = 1$. Now suppose that $(\coprod_{i \in I} A_i^{(n)})_{n \in \mathbb{N}}$ is a sequence of pairwise disjoint sets in $\bigoplus_{i \in I} \mathcal{A}_i$. Then

$$\begin{aligned}
&(\bigoplus_{i \in I} \mu_i)(x)(\bigcup_{n \in \mathbb{N}} \coprod_{i \in I} A_i^{(n)}) \\
&= (\bigoplus_{i \in I} \mu_i)(x)(\coprod_{i \in I} \bigcup_{n \in \mathbb{N}} A_i^{(n)}) \\
&= \sum_{i \in I} \mu_i(x)(\bigcup_{n \in \mathbb{N}} A_i^{(n)})
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in I}\sum_{n\in\mathbb{N}} \mu_i(x)(A_i^{(n)}) \\
&= \sum_{n\in\mathbb{N}}\sum_{i\in I} \mu_i(x)(A_i^{(n)}) \\
&= \sum_{n\in\mathbb{N}} (\bigoplus_{i\in I} \mu_i)(x)(\coprod_{i\in I} A_i^{(n)}).
\end{aligned}
$$

Thus, for all $x \in X$, $(\bigoplus_{i\in I} \mu_i)(x)$ is a probability measure on $(\coprod_{i\in I} X_i, \bigoplus_{i\in I} \mathcal{A}_i)$ and so $\bigoplus_{i\in I} \mu_i$ is well-defined.

Finally, for all $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$, $\lambda x.(\bigoplus_{i\in I} \mu_i)(x)(\coprod_{i\in I} A_i) : X \to \mathbb{R}$ is measurable because each μ_i is measurable. Thus, by Proposition A.2.4, $\bigoplus_{i\in I} \mu_i$ is a probability kernel. □

Now the converse problem of deconstructing a probability kernel $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ into its addenda is considered.

Proposition A.9.2. *Let $(X, \mathcal{A})$ a measurable space, $(X_i, \mathcal{A}_i)$ be a measurable space, for all $i \in I$, and $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ a probability kernel. Then there exists a probability kernel $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$, for all $i \in I$, such that $\sum_{i\in I} \mu_i(x)(X_i) = 1$, for all $x \in X$, and $\mu = \bigoplus_{i\in I} \mu_i$.*

Proof. For all $i \in I$, define $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ by

$$
\begin{aligned}
&\mu_i(x)(A_i) = \mu(x)(A_i) \\
&\mu_i(x)(A_i \sqcup \{*\}) = \mu(x)(A_i) + \mu(x)(\coprod_{j\in I, j\neq i} X_j)
\end{aligned}
$$

for all $x \in X$ and $A_i \in \mathcal{A}_i$. Note that $\sum_{i\in I} \mu_i(x)(X_i) = 1$, for all $x \in X$. Clearly, each μ_i is well-defined and a probability kernel. Also, for all $x \in X$ and $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$, $\mu(x)(\coprod_{i\in I} A_i) = \sum_{i\in I} \mu_i(x)(A_i)$. Thus $\mu = \bigoplus_{i\in I} \mu_i$. □

The proof of Proposition A.9.2 shows why it is natural to use the signature $X \to \mathcal{P}(X_i \sqcup \{*\})$ instead of $X \to \mathcal{P}(X_i)$ for each μ_i in Definition A.9.1.

The next result establishes that if each μ_i is a regular conditional distribution, then so is $\bigoplus_{i\in I} \mu_i$.

Proposition A.9.3. *Let $(\Omega, \mathcal{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ a measurable space, and $(X_i, \mathcal{A}_i)$ a measurable space, for all $i \in I$. Suppose that $f : \Omega \to X$ and, for all $i \in I$, $g_i : \Omega \to X_i \sqcup \{*\}$ are measurable, and, for all $i \in I$, $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ is a probability kernel that is a regular conditional distribution of g_i given f. Suppose also that $\sum_{i\in I} \mu_i(x)(X_i) = 1$, for all $x \in X$, and $(g_i^{-1}(X_i))_{i\in I}$ is a partition of Ω. Then there is a measurable function $g : \Omega \to \coprod_{i\in I} X_i$ such that $\bigoplus_{i\in I} \mu_i : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ is a regular conditional distribution of g given f.*

Proof. The probability kernel $\bigoplus_{i\in I} \mu_i : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ exists, by Proposition A.9.1. Define $g : \Omega \to \coprod_{i\in I} X_i$ by $g(\omega) = g_i(\omega)$, whenever $\omega \in g_i^{-1}(X_i)$, for all $\omega \in \Omega$. Then g is well-defined and measurable, since $(g_i^{-1}(X_i))_{i\in I}$ is a partition of Ω. It remains to show that $\bigoplus_{i\in I} \mu_i$ is a regular conditional distribution of g given f.

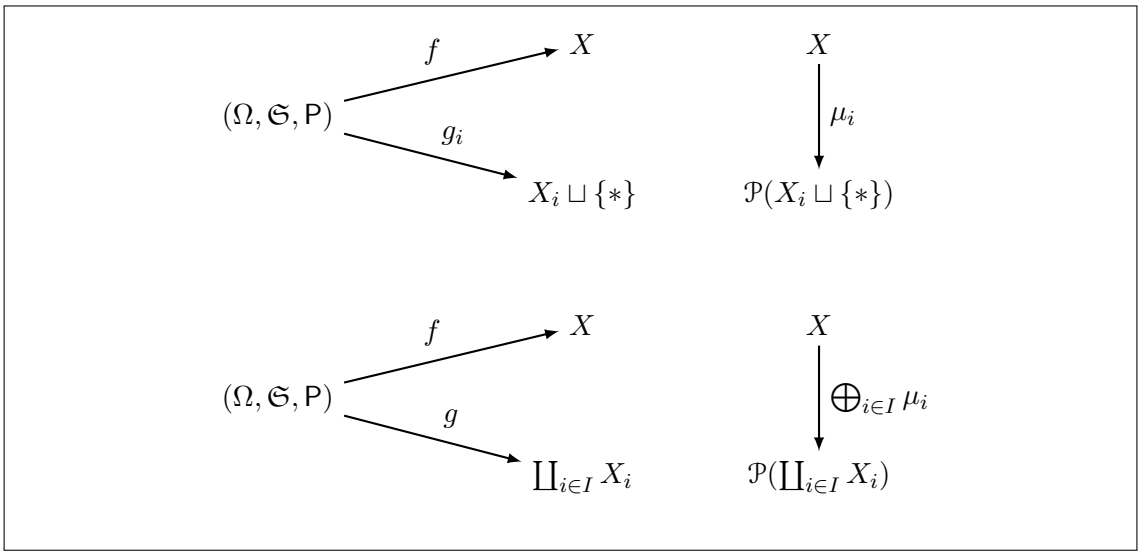

Figure A.35: Setting for Proposition A.9.3

Suppose that I is countably infinite, so that it can be assumed that $I = \mathbb{N}$. Then, for all $\coprod_{n \in \mathbb{N}} A_n \in \bigoplus_{n \in \mathbb{N}} \mathcal{A}_n$, almost surely,

$$
\begin{aligned}
& \mathsf{P}(g^{-1}(\coprod_{n \in \mathbb{N}} A_n) \mid f) \\
= \; & \mathsf{P}(\bigcup_{n \in \mathbb{N}} g^{-1}(A_n) \mid f) \\
= \; & \mathsf{E}(\mathbf{1}_{\bigcup_{n \in \mathbb{N}} g^{-1}(A_n)} \mid f) \\
= \; & \mathsf{E}(\sum_{n \in \mathbb{N}} \mathbf{1}_{g^{-1}(A_n)} \mid f) && [(g^{-1}(A_n))_{n \in \mathbb{N}} \text{ are pairwise disjoint}] \\
= \; & \mathsf{E}(\lim_{n \to \infty} \sum_{i=1}^{n} \mathbf{1}_{g^{-1}(A_i)} \mid f) \\
= \; & \lim_{n \to \infty} \mathsf{E}(\sum_{i=1}^{n} \mathbf{1}_{g^{-1}(A_i)} \mid f) && [\text{Proposition A.5.6, Part 4}] \\
= \; & \lim_{n \to \infty} \sum_{i=1}^{n} \mathsf{E}(\mathbf{1}_{g^{-1}(A_i)} \mid f) && [\text{Proposition A.5.6, Part 1}] \\
= \; & \sum_{n \in \mathbb{N}} \mathsf{E}(\mathbf{1}_{g^{-1}(A_n)} \mid f) \\
= \; & \sum_{n \in \mathbb{N}} \mathsf{P}(g^{-1}(A_n) \mid f) \\
= \; & \sum_{n \in \mathbb{N}} \mathsf{P}(g_n^{-1}(A_n) \mid f) \\
= \; & \sum_{n \in \mathbb{N}} \lambda\omega.\mu_n(f(\omega))(A_n) && [\text{Each } \mu_n \text{ is a regular conditional distribution}]
\end{aligned}
$$

$$= \lambda\omega.(\bigoplus_{n\in\mathbb{N}} \mu_n)(f(\omega))(\coprod_{n\in\mathbb{N}} A_n).$$

The case when I is finite is similar. Thus $\bigoplus_{i\in I} \mu_i$ is a regular conditional distribution of g given f. □

Proposition A.9.2 shows that the assumption that $\sum_{i\in I} \mu_i(x)(X_i) = 1$, for all $x \in X$, in Proposition A.9.3 is necessary. Also, Proposition A.9.4 below shows that the assumption that $(g_i^{-1}(X_i))_{i\in I}$ is a partition of Ω in Proposition A.9.3 is necessary.

Another reason for the introduction of the one point extensions in Proposition A.9.3 should be clear from the proof. If each g_i has signature $\Omega \to X_i$, there is no reasonable way of defining g. Instead, each g_i has signature $\Omega \to X_i \sqcup \{*\}$ to allow the points in Ω which should map to some other X_j $(j \neq i)$ to be mapped to $*$.

The next result establishes that if $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ is a regular conditional distribution, then $\mu = \bigoplus_{i\in I} \mu_i$, where each μ_i is a regular conditional distribution. This is a converse to Proposition A.9.3.

Proposition A.9.4. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ and, for all $i \in I$, $(X_i, \mathcal{A}_i)$ a measurable space, $f : \Omega \to X$ and $g : \Omega \to \coprod_{i\in I} X_i$ measurable functions, and $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ a probability kernel that is a regular conditional distribution of g given f. Then there exists a measurable function $g_i : \Omega \to X_i \sqcup \{*\}$ and a probability kernel $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ that is a regular conditional distribution of g_i given f, for all $i \in I$, such that $(g_i^{-1}(X_i))_{i\in I}$ is a partition of Ω and $\mu = \bigoplus_{i\in I} \mu_i$.*

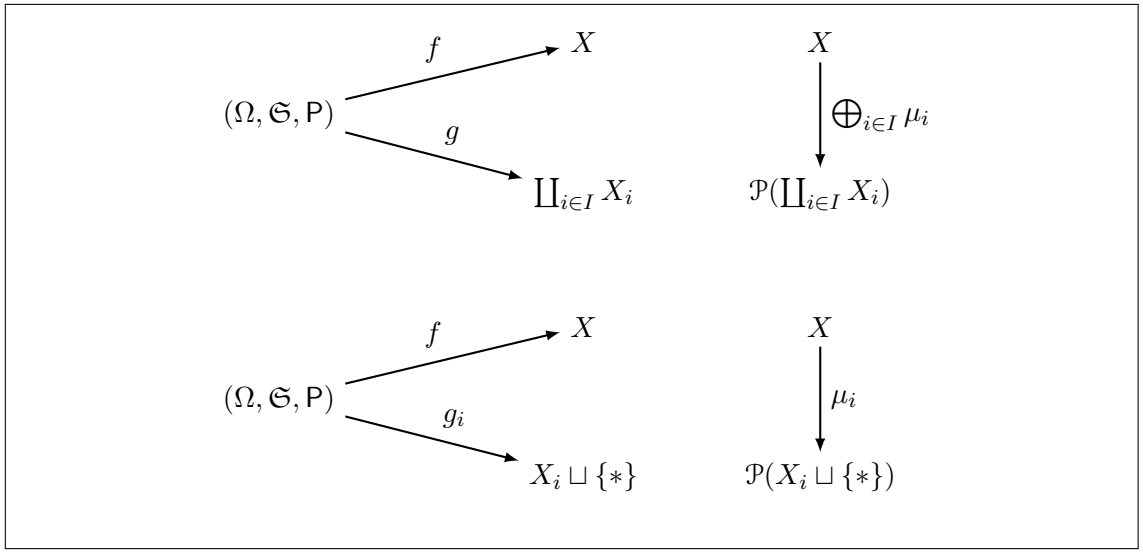

Figure A.36: Setting for Proposition A.9.4

Proof. By Proposition A.9.2, there exists a probability kernel $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$, for all $i \in I$ such that $\mu = \bigoplus_{i\in I} \mu_i$.

For all $i \in I$, define $g_i : \Omega \to X_i \sqcup \{*\}$ by

$$g_i(\omega) = \begin{cases} g(\omega) & \text{if } \omega \in g^{-1}(X_i) \\ * & \text{otherwise,} \end{cases}$$

for all $\omega \in \Omega$. Clearly, each g_i is measurable and $(g_i^{-1}(X_i))_{i\in I}$ is a partition of Ω.

It remains to show that each μ_i is a regular conditional distribution of g_i given f. For all $i \in I$ and $A_i \in \mathcal{A}_i$, almost surely,

$$\begin{aligned}
&\mathsf{P}(g_i^{-1}(A_i) \mid f) \\
&= \mathsf{P}(g^{-1}(A_i) \mid f) \\
&= \lambda\omega.\mu(f(\omega))(A_i) && [\mu \text{ is a regular conditional distribution}] \\
&= \lambda\omega.\mu_i(f(\omega))(A_i).
\end{aligned}$$

Also, for all $i \in I$ and $A_i \in \mathcal{A}_i$, almost surely,

$$\begin{aligned}
&\mathsf{P}(g_i^{-1}(A_i \sqcup \{*\}) \mid f) \\
&= \mathsf{P}(g^{-1}(A_i) \cup g^{-1}(\coprod_{j\in I, j\neq i} X_j) \mid f) \\
&= \lambda\omega.(\mu(f(\omega))(A_i) + \mu(f(\omega))(\coprod_{j\in I, j\neq i} X_j)) && [\mu \text{ is a regular conditional distribution}] \\
&= \lambda\omega.(\mu_i(f(\omega))(A_i) + \mu_i(f(\omega))(\{*\})) && [\text{Proof of Proposition A.9.2}] \\
&= \lambda\omega.\mu_i(f(\omega))(A_i \sqcup \{*\}).
\end{aligned}$$

It follows that each μ_i is a regular conditional distribution of g_i given f. □

Proposition A.9.5. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and, for all $i \in I$, $(Z_i, \mathcal{C}_i)$ measurable spaces and $\mu_i : X \times Y \to \mathcal{P}(Z_i \sqcup \{*\})$ probability kernels. Suppose that $\sum_{i\in I} \mu_i(x, y)(Z_i) = 1$, for all $x \in X$ and $y \in Y$. Then, for all $x \in X$,*

$$\lambda y.(\bigoplus_{i\in I} \mu_i)(x, y) = \bigoplus_{i\in I} \lambda y.\mu_i(x, y).$$

Proof. Note that $\bigoplus_{i\in I} \mu_i : X\times Y \to \mathcal{P}(\coprod_{i\in I} Z_i)$ is well-defined since $\sum_{i\in I} \mu_i(x, y)(Z_i) = 1$, for all $x \in X$ and $y \in Y$. By Proposition A.2.5, for all $x \in X$, $\lambda y.(\bigoplus_{i\in I} \mu_i)(x, y) : Y \to \mathcal{P}(\coprod_{i\in I} Z_i)$ is a probability kernel.

Also, by Proposition A.2.5, for all $i \in I$ and $x \in X$, $\lambda y.\mu_i(x, y) : Y \to \mathcal{P}(Z_i \sqcup \{*\})$ is a probability kernel. Then, for all $x \in X$, $\bigoplus_{i\in I} \lambda y.\mu_i(x, y) : Y \to \mathcal{P}(\coprod_{i\in I} Z_i)$ is well-defined since $\sum_{i\in I} \mu_i(x, y)(Z_i) = 1$, for all $y \in Y$.

Now, for all $x \in X$, $y \in Y$, and $\coprod_{i\in I} C_i \in \bigoplus_{i\in I} \mathcal{C}_i$,

$$\begin{aligned}
&(\lambda y.(\bigoplus_{i\in I} \mu_i)(x, y))(y)(\coprod_{i\in I} C_i) \\
&= (\bigoplus_{i\in I} \mu_i)(x, y)(\coprod_{i\in I} C_i) \\
&= \sum_{i\in I} \mu_i(x, y)(C_i) \\
&= \sum_{i\in I} \lambda y.\mu_i(x, y)(y)(C_i) \\
&= (\bigoplus_{i\in I} \lambda y.\mu_i(x, y))(y)(\coprod_{i\in I} C_i).
\end{aligned}$$

Hence the result. □

The next result is used for computing integrals with respect to sums of probability kernels.

Proposition A.9.6. *Let $(X, \mathcal{A})$ be a measurable space and, for all $i \in I$, $(X_i, \mathcal{A}_i)$ a measurable space and $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ a probability kernel. Suppose that $\sum_{i \in I} \mu_i(x)(X_i) = 1$, for all $x \in X$. Let $f : \coprod_{i \in I} X_i \to \mathbb{R}$ be a non-negative measurable function. Then, for all $x \in X$,*

$$\int_{\coprod_{i \in I} X_i} f \; d(\bigoplus_{i \in I} \mu_i)(x) = \sum_{i \in I} \int_{X_i} f|_{X_i} \; d\mu_i(x)|_{\mathcal{A}_i}.$$

Proof. The result holds when f is a measurable indicator function, and hence holds for simple functions. Now apply the monotone convergence theorem (Proposition A.2.2). □

By taking X to be a singleton set, the definition of the sum of probability kernels reduces to the definition of the sum of probability measures, as follows. Let $(X_i, \mathcal{A}_i)$ be a measurable space and $\mu_i : \mathcal{P}(X_i \sqcup \{*\})$ a probability measure, for all $i \in I$. Suppose that $\sum_{i \in I} \mu_i(X_i) = 1$. Then the sum $\bigoplus_{i \in I} \mu_i : \mathcal{P}(\coprod_{i \in I} X_i)$ of $(\mu_i)_{i \in I}$ is defined by

$$(\bigoplus_{i \in I} \mu_i)(\coprod_{i \in I} A_i) = \sum_{i \in I} \mu_i(A_i),$$

for all $\coprod_{i \in I} A_i \in \bigoplus_{i \in I} \mathcal{A}_i$.

To sample from a probability space $(\coprod_{i \in I} X_i, \bigoplus_{i \in I} \mathcal{A}_i, \bigoplus_{i \in I} \mu_i)$, first sample from $(I, \mathbb{B}^I, \lambda i.\mu_i(X_i) \cdot c)$ to produce a $j \in I$, then sample from $(X_j, \mathcal{A}_j, \mu_j|_{X_j})$ to produce an $x \in X_j$. (Recall that c is counting measure.)

While the approach taken so far to sums has the desirable characteristic of preserving the property of being a regular conditional distribution (Propositions A.9.3 and A.9.4), it turns out to be problematical for the intended applications of this book. First, Proposition A.9.4 demands the introduction of the set $\{*\}$ into the codomain of each $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$. This means that further deconstruction of each μ_i is obstructed by the presence of $\{*\}$. Given that, in applications, it will nearly always be desirable to deconstruct each μ_i, this is a problem. Second, Proposition A.9.3 has the requirement that $\sum_{i \in I} \mu_i(x)(X_i) = 1$, for all $x \in X$, in order to form $\bigoplus_{i \in I} \mu_i$, which means that the filtering of each of the μ_i cannot be done independently.

In view of these problems, an alternative way of thinking about a sum of probability kernels is desirable. The next proposition suggests a way forward.

Proposition A.9.7. *Let $(X, \mathcal{A})$ be a measurable space and, for all $i \in I$, $(X_i, \mathcal{A}_i)$ a measurable space.*

1. *For all $i \in I$, let $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ be a probability kernel such that $\mu_i(x)(X_i) > 0$, for all $x \in X$. Suppose also that $\sum_{i \in I} \mu_i(x)(X_i) = 1$, for all $x \in X$. Then, for all $i \in I$, there exists a probability kernel $\overline{\mu}_i : X \to \mathcal{P}(X_i)$ and a measurable function $h_i : X \to \mathbb{R}$, where $h_i > 0$ and $\sum_{i \in I} h_i(x) = 1$, for all $x \in X$, such that*

$$(\bigoplus_{i \in I} \mu_i)(x)(\coprod_{i \in I} A_i) = \sum_{i \in I} h_i(x) \, \overline{\mu}_i(x)(A_i),$$

for all $x \in X$ and $\coprod_{i \in I} A_i \in \bigoplus_{i \in I} \mathcal{A}_i$.

2. *For all $i \in I$, let $\overline{\mu}_i : X \to \mathcal{P}(X_i)$ be a probability kernel and $h_i : X \to \mathbb{R}$ a measurable function such that $h_i > 0$ and $\sum_{i\in I} h_i(x) = 1$, for all $x \in X$. Then, for all $i \in I$, there exists a probability kernel $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ such that $\sum_{i\in I} \mu_i(x)(X_i) = 1$, for all $x \in X$, and*

$$(\bigoplus_{i\in I} \mu_i)(x)(\coprod_{i\in I} A_i) = \sum_{i\in I} h_i(x)\, \overline{\mu}_i(x)(A_i),$$

for all $x \in X$ and $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$.

Proof. 1. For all $i \in I$, define $\overline{\mu}_i : X \to \mathcal{P}(X_i)$ by

$$\overline{\mu}_i(x)(A_i) = \frac{\mu_i(x)(A_i)}{\mu_i(x)(X_i)},$$

for all $x \in X$ and $A_i \in \mathcal{A}_i$. Also, for all $i \in I$, define $h_i : X \to \mathbb{R}$ by

$$h_i(x) = \mu_i(x)(X_i),$$

for all $x \in X$. Clearly, $(\overline{\mu}_i)_{i\in I}$ and $(h_i)_{i\in I}$ have the desired properties.

2. For all $i \in I$, define $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ by

$$\mu_i(x)(A_i) = h_i(x)\, \overline{\mu}_i(x)(A_i),$$

and

$$\mu_i(x)(A_i \sqcup \{*\}) = h_i(x)\, \overline{\mu}_i(x)(A_i) + \sum_{j\in I\setminus\{i\}} h_j(x),$$

for all $x \in X$ and $A_i \in \mathcal{A}_i$. Clearly, $(\mu_i)_{i\in I}$ has the desired properties. □

Thus, to form a sum of probability kernels, instead of starting with the ingredients of Definition A.9.1, one can start with the $(\overline{\mu}_i : X \to \mathcal{P}(X_i))_{i\in I}$ and $(h_i : X \to \mathbb{R})_{i\in I}$ of Part 2 of Proposition A.9.7. This suggests the following approach to handling sums.

Let $(X_i, \mathcal{A}_i, \mu_i)$ be a probability space, for all $i \in I$. Suppose $h : I \to [0, \infty)$ is a function such that $\sum_{i\in I} h(i) = 1$. (Thus h is a density on the measure space $(I, \mathbb{B}^I, c)$, where c is counting measure.) Define

$$h \bullet \bigoplus_{i\in I} \mu_i : \bigoplus_{i\in I} \mathcal{A}_i \to [0, \infty)$$

by

$$(h \bullet \bigoplus_{i\in I} \mu_i)(\coprod_{i\in I} A_i) = \sum_{i\in I} h(i)\, \mu_i(A_i),$$

for all $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$. Then

$$\begin{aligned}
&(h \bullet \bigoplus_{i\in I} \mu_i)(\coprod_{i\in I} X_i) \\
&= \sum_{i\in I} h(i)\, \mu_i(X_i) \\
&= \sum_{i\in I} h(i) \\
&= 1.
\end{aligned}$$

Next suppose that $(\coprod_{i\in I} A_i^{(n)})_{n\in\mathbb{N}_0}$ is a sequence of pairwise disjoint sets in $\bigoplus_{i\in I}\mathcal{A}_i$. Then

$$\begin{aligned}
&(h\bullet\bigoplus_{i\in I}\mu_i)(\bigcup_{n\in\mathbb{N}_0}\coprod_{i\in I}A_i^{(n)})\\
&=(h\bullet\bigoplus_{i\in I}\mu_i)(\coprod_{i\in I}\bigcup_{n\in\mathbb{N}_0}A_i^{(n)})\\
&=\sum_{i\in I}h(i)\mu_i(\bigcup_{n\in\mathbb{N}_0}A_i^{(n)})\\
&=\sum_{i\in I}h(i)\sum_{n\in\mathbb{N}_0}\mu_i(A_i^{(n)})\\
&=\sum_{n\in\mathbb{N}_0}\sum_{i\in I}h(i)\mu_i(A_i^{(n)})\\
&=\sum_{n\in\mathbb{N}_0}(h\bullet\bigoplus_{i\in I}\mu_i)(\coprod_{i\in I}A_i^{(n)}).
\end{aligned}$$

Thus $h\bullet\bigoplus_{i\in I}\mu_i$ is a probability measure on $(\coprod_{i\in I}X_i,\bigoplus_{i\in I}\mathcal{A}_i)$.

This idea is now generalized to the concept of the weighted sum of probability kernels.

Definition A.9.2. Let I be an index set and $(X,\mathcal{A})$ a measurable space. A *weight function* is a function $h:X\to I\to[0,\infty)$ such that $\sum_{i\in I}h(x)(i)=1$, for all $x\in X$, and $\lambda x.h(x)(i):X\to[0,\infty)$ is measurable, for all $i\in I$.

Definition A.9.3. Let $(X,\mathcal{A})$ and $(X_i,\mathcal{A}_i)$, for all $i\in I$, be measurable spaces, $\mu_i:X\to\mathcal{P}(X_i)$ a probability kernel, for all $i\in I$, and $h:X\to I\to[0,\infty)$ a weight function. Define the *weighted sum* of $(\mu_i)_{i\in I}$ with respect to h

$$h\bullet\bigoplus_{i\in I}\mu_i:X\to\mathcal{P}(\coprod_{i\in I}X_i)$$

by

$$(h\bullet\bigoplus_{i\in I}\mu_i)(x)(\coprod_{i\in I}A_i)=\sum_{i\in I}h(x)(i)\,\mu_i(x)(A_i),$$

for all $x\in X$ and $\coprod_{i\in I}A_i\in\bigoplus_{i\in I}\mathcal{A}_i$. Each μ_i is called an *addendum* of the weighted sum.

Thus $(h\bullet\bigoplus_{i\in I}\mu_i)(x)=h(x)\bullet\bigoplus_{i\in I}\mu_i(x)$, for all $x\in X$. By remarks just above, $h\bullet\bigoplus_{i\in I}\mu_i$ is well-defined. Now it is shown that $h\bullet\bigoplus_{i\in I}\mu_i$ is a probability kernel.

Proposition A.9.8. *Let $(X,\mathcal{A})$ and $(X_i,\mathcal{A}_i)$, for all $i\in I$, be measurable spaces, $\mu_i:X\to\mathcal{P}(X_i)$ a probability kernel, for all $i\in I$, and $h:X\to I\to[0,\infty)$ a weight function. Then $h\bullet\bigoplus_{i\in I}\mu_i$ is a probability kernel.*

Proof. For $\coprod_{i\in I}A_i\in\bigoplus_{i\in I}\mathcal{A}_i$, consider

$$\lambda x.(h\bullet\bigoplus_{i\in I}\mu_i)(x)(\coprod_{i\in I}A_i):X\to\mathbb{R}.$$

Now $\lambda x.(h\bullet\bigoplus_{i\in I}\mu_i)(x)(\coprod_{i\in I}A_i)=\lambda x.\sum_{i\in I}h(x)(i)\,\mu_i(x)(A_i)$, which is measurable being a sum of measurable real-valued functions. By Proposition A.2.4, it follows that $h\bullet\bigoplus_{i\in I}\mu_i$ is measurable. Thus $h\bullet\bigoplus_{i\in I}\mu_i$ is a probability kernel. □

In the converse direction, the next proposition shows that a probability kernel which maps into a space of probability measures on sum spaces can be deconstructed.

Proposition A.9.9. *Let $(X, \mathcal{A})$ and $(X_i, \mathcal{A}_i)$, for all $i \in I$, be measurable spaces, and $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ a probability kernel. Suppose that $\mu(x)(X_i) > 0$, for all $x \in X$ and $i \in I$. Then there exists a weight function $h : X \to I \to [0, \infty)$ and, for all $i \in I$, a probability kernel $\mu_i : X \to \mathcal{P}(X_i)$ such that $\mu = h \bullet \bigoplus_{i\in I} \mu_i$.*

Proof. Define $h : X \to I \to [0, \infty)$ by $h(x)(i) = \mu(x)(X_i)$, for all $x \in X$ and $i \in I$. Clearly, $\lambda x.h(x)(i) : X \to [0, \infty)$ is measurable, for all $i \in I$, and $\sum_{i\in I} h(x)(i) = 1$, for all $x \in X$. For $i \in I$, define $\mu_i : X \to \mathcal{P}(X_i)$ by $\mu_i(x)(A_i) = \mu(x)(A_i)/\mu(x)(X_i)$, for all $x \in X$ and $A_i \in \mathcal{A}_i$. Then each μ_i and $h \bullet \bigoplus_{i\in I} \mu_i : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ is a probability kernel. Furthermore, for all $x \in X$ and $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$,

$$\begin{aligned}
&(h \bullet \bigoplus_{i\in I} \mu_i)(x)(\coprod_{i\in I} A_i) \\
&= \sum_{i\in I} h(x)(i)\, \mu_i(x)(A_i) \\
&= \sum_{i\in I} \mu(x)(X_i)\, (\mu(x)(A_i)/\mu(x)(X_i)) \\
&= \sum_{i\in I} \mu(x)(A_i) \\
&= \mu(x)(\coprod_{i\in I} A_i).
\end{aligned}$$

Hence $\mu = h \bullet \bigoplus_{i\in I} \mu_i$. □

It will be crucial to establish that $h \bullet \bigoplus_{i\in I} \mu_i$ is a regular conditional distribution if each μ_i is a regular conditional distribution. Here is the setting for this.

Let $(X, \mathcal{A})$ be measurable space and $(I, \mathbb{B}^I, c)$ a countable measure space. Suppose that $f : \Omega \to X$ and $e : \Omega \to I$ are measurable. Let $\chi : X \to \mathcal{P}(I)$ be a regular conditional distribution of e given f. Thus

$$\mathsf{P}(e^{-1}(J) \mid f) = \lambda\omega.\chi(f(\omega))(J) \text{ a.s.},$$

for all $J \in \mathbb{B}^I$. Note that $\chi = \breve{\chi} \cdot c$, where $\breve{\chi} : X \to \mathcal{D}(I)$ is the conditional density defined by $\breve{\chi}(x)(i) = \chi(x)(\{i\})$, for all $x \in X$ and $i \in I$. Also $\breve{\chi}$ is a weight function.

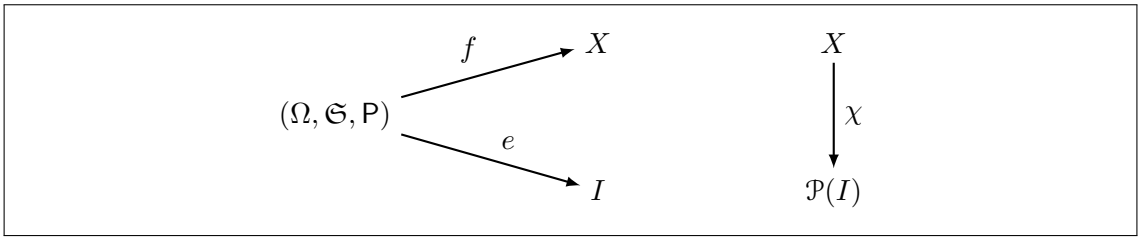

Figure A.37: Setting for the probability kernel χ

Continuing the setting, let $(X_i, \mathcal{A}_i)$ be a measurable space and $g_i : \Omega \to X_i$ measurable function, for all $i \in I$. Suppose that, for all $i \in I$, $\mu_i : X \to \mathcal{P}(X_i)$ is a regular condition distribution of g_i given f. Thus, for all $i \in I$,

$$\mathsf{P}(g_i^{-1}(A_i) \mid f) = \lambda\omega.\mu_i(f(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$.

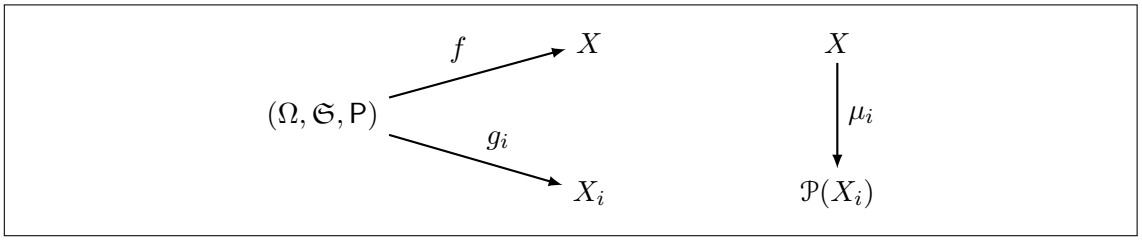

Figure A.38: Setting for the μ_i

Now define $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ by

$$\mu = \breve{\chi} \bullet \bigoplus_{i\in I} \mu_i.$$

Thus

$$\mu(x)(\coprod_{i\in I} A_i) = \sum_{i\in I} \breve{\chi}(x)(i)\, \mu_i(x)(A_i),$$

for all $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$. It was shown above that μ is a probability kernel. Under a natural conditional independence condition, the next proposition shows that μ is a regular conditional distribution. This provides an ideal approach to dealing with schemas whose codomain is probability measures on a sum space: the awkward $\{*\}$ needed earlier is gone, μ is a regular conditional distribution which is needed for schemas, and the μ_i are regular conditional distributions and can easily be deconstructed further, if necessary.

Proposition A.9.10. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ and $(X_i, \mathcal{A}_i)$, for all $i \in I$, measurable spaces, and $(I, \mathbb{B}^I, c)$ a countable measure space. Suppose that $f : \Omega \to X$, $e : \Omega \to I$, and $g_i : \Omega \to X_i$, for all $i \in I$, are measurable functions. Suppose also that, for all $i \in I$, $\mu_i : X \to \mathcal{P}(X_i)$ is a regular condition distribution of g_i given f and $\chi : X \to \mathcal{P}(I)$ is a regular conditional distribution of e given f. Define $\breve{\chi} : X \to \mathcal{D}(I)$ by $\breve{\chi}(x)(i) = \chi(x)(\{i\})$, for all $x \in X$ and $i \in I$, $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ by $\mu = \breve{\chi} \bullet \bigoplus_{i\in I} \mu_i$, and $g : \Omega \to \coprod_{i\in I} X_i$ by $g = \lambda\omega.g_{e(\omega)}(\omega)$. Suppose that $\sigma(e)$ and $\sigma(g_i)$ are conditionally independent given $\sigma(f)$, for all $i \in I$. Then μ is a regular conditional distribution of g given f.*

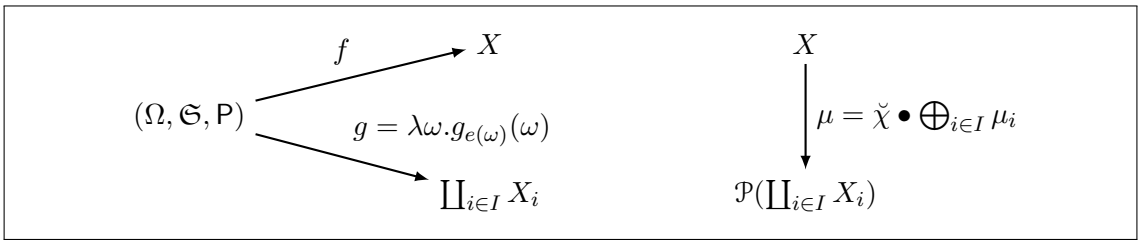

Figure A.39: Setting for μ in Proposition A.9.10

Proof. First, it is shown that g is measurable. For this, let $A_i \in \mathcal{A}_i$. Then

$$\begin{aligned}
& g^{-1}(A_i) \\
&= \{\omega \mid g(\omega) \in A_i\} \\
&= \{\omega \mid e(\omega) = i \text{ and } g_i(\omega) \in A_i\} \\
&= \{\omega \mid e(\omega) = i\} \cap \{\omega \mid g_i(\omega) \in A_i\} \\
&= e^{-1}(\{i\}) \cap g_i^{-1}(A_i).
\end{aligned}$$

Thus, for all $i \in I$, $g^{-1}(A_i) = e^{-1}(\{i\}) \cap g_i^{-1}(A_i)$. Since e and each g_i are measurable, $g^{-1}(A_i)$ is measurable. Now let $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$. Then $g^{-1}(\coprod_{i\in I} A_i) = \bigcup_{i\in I} g^{-1}(A_i)$, being the countable union of measurable sets is measurable. Hence g is measurable.

Next it is shown that $\breve{\chi}$ is a conditional density. For this, it needs to be shown that $\lambda(x,i).\breve{\chi}(x)(i) : X \times I \to \mathbb{R}$ is measurable. For all $i \in I$, the function $\lambda x.\breve{\chi}(x)(i) : X \to \mathbb{R}$ is measurable, since χ is measurable. Thus $\lambda(x,i).\breve{\chi}(x)(i)$ is separately measurable on X. Now give I the discrete topology. Then I is a separable metrizable space. Also $\mathbb{B}^I$ is (trivially) the Borel σ-algebra generated by the discrete topology and $\mathbb{R}$ is a metrizable space. Finally, note that, for all $x \in X$, $\lambda i.\breve{\chi}(x)(i) : I \to \mathbb{R}$ is continuous, since I has the discrete topology. Thus $\lambda(x,i).\breve{\chi}(x)(i)$ is separately continuous on I. Separate measurability on X, separate continuity on I, I being a separable metrizable space, and $\mathbb{R}$ being a metrizable space are the conditions of Proposition A.4.4 needed to establish that $\lambda(x,i).\breve{\chi}(x)(i)$ is (jointly) measurable.

Suppose that I is countably infinite, so that it can be assumed that $I = \mathbb{N}$. Then, for all $\coprod_{n\in\mathbb{N}} A_n \in \bigoplus_{n\in\mathbb{N}} \mathcal{A}_n$, almost surely,

$$\begin{aligned}
& \mathsf{P}(g^{-1}(\coprod_{n\in\mathbb{N}} A_n) \mid f) \\
&= \mathsf{P}(\bigcup_{n\in\mathbb{N}} g^{-1}(A_n) \mid f) \\
&= \mathsf{E}(\mathbf{1}_{\bigcup_{n\in\mathbb{N}} g^{-1}(A_n)} \mid f) \\
&= \mathsf{E}(\sum_{n\in\mathbb{N}} \mathbf{1}_{g^{-1}(A_n)} \mid f) && [(g^{-1}(A_n))_{n\in\mathbb{N}} \text{ are pairwise disjoint}] \\
&= \mathsf{E}(\lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{1}_{g^{-1}(A_i)} \mid f) \\
&= \lim_{n\to\infty} \mathsf{E}(\sum_{i=1}^{n} \mathbf{1}_{g^{-1}(A_i)} \mid f) && [\text{Proposition A.5.6, Part 4}]
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mathsf{E}(\mathbf{1}_{g^{-1}(A_i)} \mid f) && \text{[Proposition A.5.6, Part 1]}\\
&= \sum_{n\in\mathbb{N}} \mathsf{E}(\mathbf{1}_{g^{-1}(A_n)} \mid f)\\
&= \sum_{n\in\mathbb{N}} \mathsf{P}(g^{-1}(A_n) \mid f)\\
&= \sum_{n\in\mathbb{N}} \mathsf{P}(e^{-1}(\{n\}) \cap g_n^{-1}(A_n) \mid f)\\
&= \sum_{n\in\mathbb{N}} \mathsf{P}(e^{-1}(\{n\}) \mid f)\ \mathsf{P}(g_n^{-1}(A_n) \mid f) && [\sigma(e) \underset{\sigma(f)}{\perp\!\!\!\perp} \sigma(g_n), \text{ for all } n \in \mathbb{N}]\\
&= \sum_{n\in\mathbb{N}} \lambda\omega.\breve{\chi}(f(\omega))(n)\ \mu_n(f(\omega))(A_n) && [\chi \text{ and each } \mu_n \text{ is a reg. cond. distr.}]\\
&= \lambda\omega.\mu(f(\omega))(\coprod_{n\in\mathbb{N}} A_n).
\end{aligned}
$$

The case when I is finite is similar. Thus $\breve{\chi} \bullet \bigoplus_{i\in I} \mu_i$ is a regular conditional distribution of g given f. □

Is there a converse to Proposition A.9.10? Such a result would state that a regular conditional distribution $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ could be deconstructed into regular conditional distributions $\mu_i : X \to \mathcal{P}(X_i)$, for all $i \in I$, such that $\mu = \bigoplus_{i\in I} \mu_i$ or, perhaps, $\mu = \breve{\chi} \bullet \bigoplus_{i\in I} \mu_i$, for some χ. The first possibility, that $\mu = \bigoplus_{i\in I} \mu_i$, seems out of the question for arbitrary μ. One serious difficulty is defining the g_i from g. This was exactly why $\{*\}$ was introduced earlier in Proposition A.9.4 and elsewhere, and leads to the μ_i having signature $X \to \mathcal{P}(X_i \sqcup \{*\})$, with the deconstruction problems that the presence of the $\{*\}$ entails.

For the second possibility, that $\mu = \breve{\chi} \bullet \bigoplus_{i\in I} \mu_i$, Proposition A.9.9 shows how to define χ and each μ_i. The function e could be defined by $e(\omega) = j$, where $g(\omega) \in X_j$, for all $\omega \in \Omega$. But the problem once again is to define each g_i from g. It seems that one effectively has to assume that there exist g_i such that g has the form $\lambda\omega.g_{e(\omega)}(\omega)$. This is such a strong assumption on g that it makes the converse trivial. The class of regular conditional distributions with signature $X \to \mathcal{P}(\coprod_{i\in I} X_i)$ having the form $\breve{\chi} \bullet \bigoplus_{i\in I} \mu_i$ is a subclass of the class of all regular conditional distributions with signature $X \to \mathcal{P}(\coprod_{i\in I} X_i)$. But it seems to provide a rich subclass of probability kernels that would be useful for the majority of applications. The strategy will therefore be to stick to this subclass which also has the significant advantage of being easily deconstructed, by definition. A moral of this tale is that, while probability kernels for the product case are straightforward, probability kernels for the sum case are problematical.

The next result is used for computing integrals with respect to sums of probability kernels.

Proposition A.9.11. *Let $(X, \mathcal{A})$ and $(X_i, \mathcal{A}_i)$, for all $i \in I$, be measurable spaces, and $h \bullet \bigoplus_{i\in I} \mu_i : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ a probability kernel, where $h : X \to I \to [0, \infty)$ is a weight function and, for all $i \in I$, $\mu_i : X \to \mathcal{P}(X_i)$ is a probability kernel. Suppose that*

$f : \coprod_{i\in I} X_i \to \mathbb{R}$ is a non-negative measurable function. Then, for all $x \in X$,

$$\int_{\coprod_{i\in I} X_i} f \; d(h \bullet \bigoplus_{i\in I} \mu_i)(x) = \sum_{i\in I} \int_{X_i} h(x)(i) f|_{X_i} \; d\mu_i(x).$$

Proof. The result holds when f is a measurable indicator function, and hence holds for simple functions. Now apply the monotone convergence theorem (Proposition A.2.2). □

To sample from a probability space $(\coprod_{i\in I} X_i, \bigoplus_{i\in I} \mathcal{A}_i, h \bullet \bigoplus_{i\in I} \mu_i)$, first sample from $(I, \mathbb{B}^I, h \cdot c)$ to produce a $j \in I$, then sample from $(X_j, \mathcal{A}_j, \mu_j)$ to produce an x_j.

Proposition A.9.12. *Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$ and $(Z_i, \mathcal{C}_i)$, for $i \in I$, be measurable spaces, $h : X \times Y \to I \to [0, \infty)$ a weight function, and $\mu_i : X \times Y \to \mathcal{P}(Z_i)$, for $i \in I$, probability kernels. Then, for all $x \in X$,*

$$\lambda y.(h \bullet \bigoplus_{i\in I} \mu_i)(x, y) = \lambda y.h(x, y) \bullet \bigoplus_{i\in I} \lambda y.\mu_i(x, y).$$

Proof. Note that, for all $x \in X$, $\lambda y.(h \bullet \bigoplus_{i\in I} \mu_i)(x, y) : Y \to \mathcal{P}(\coprod_{i\in I} Z_i)$ and, for all $i \in I$, $\lambda y.\mu_i(x, y) : Y \to \mathcal{P}(Z_i)$ are probability kernels, by Proposition A.2.5.

Now, for all $x \in X$, $y \in Y$, and $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$,

$$\begin{aligned}
&\lambda y.(h \bullet \bigoplus_{i\in I} \mu_i)(x, y)(y)(\coprod_{i\in I} A_i) \\
&= \sum_{i\in I} h(x, y)(i) \, \mu_i(x, y)(A_i) \\
&= \sum_{i\in I} h(x, y)(i) \, \lambda y.\mu_i(x, y)(y)(A_i) \\
&= (\lambda y.h(x, y) \bullet \bigoplus_{i\in I} \lambda y.\mu_i(x, y))(y)(\coprod_{i\in I} A_i).
\end{aligned}$$

Hence the result. □

Here is a connection between sum measures and projective products. It shows that, given a probability measure $\mu : \mathcal{P}(\coprod_{i\in I} X_i)$ such that $\mu = h \bullet \bigoplus_{i\in I} \mu_i$ and a non-negative integrable function $f : \coprod_{i\in I} X_i \to \mathbb{R}$, there exists a function $\overline{h}$ and, for all $i \in I$, a non-negative integrable function $h_i : X_i \to \mathbb{R}$ such that $f * \mu = \overline{h} \bullet \bigoplus_{i\in I} (h_i * \mu_i)$, and gives explicit forms for $\overline{h}$ and each h_i.

Proposition A.9.13. *Let $(X_i, \mathcal{A}_i)$ be a measurable space, for all $i \in I$, and $\mu : \mathcal{P}(\coprod_{i\in I} X_i)$ a probability measure such that $\mu(X_i) > 0$, for all $i \in I$. Suppose that $\mu = h \bullet \bigoplus_{i\in I} \mu_i$, for some function $h : I \to [0, \infty)$ and $\mu_i : \mathcal{P}(X_i)$, for all $i \in I$. Let $f : \coprod_{i\in I} X_i \to \mathbb{R}$ be a non-negative measurable function such that $0 < \int_{\coprod_{i\in I} X_i} f \; d(h \bullet \bigoplus_{i\in I} \mu_i) < \infty$ and $0 < \int_{X_i} f|_{X_i} \; d\mu_i < \infty$, for all $i \in I$. Define $h_i : X_i \to \mathbb{R}$ by $h_i = f|_{X_i}$, for all $i \in I$. Also define $\overline{h} : I \to [0, \infty)$ by*

$$\overline{h}(i) = \frac{\int_{X_i} h(i) \, f|_{X_i} \, d\mu_i}{\sum_{i\in I} \int_{X_i} h(i) \, f|_{X_i} \, d\mu_i},$$

for all $i \in I$. Then

$$f * \mu = \overline{h} \bullet \bigoplus_{i \in I} (h_i * \mu_i).$$

Proof. Note that $\sum_{i \in I} \overline{h}(i) = 1$. By Proposition A.9.9, there exists a function $h : I \to [0, \infty)$ such that $\sum_{i \in I} h(i) = 1$, and $\mu_i \in \mathcal{P}(X_i)$, for all $i \in I$, such that $\mu = h \bullet \bigoplus_{i \in I} \mu_i$.

For all $\coprod_{i \in I} A_i \in \bigoplus_{i \in I} \mathcal{A}_i$,

$$\begin{aligned}
&(\overline{h} \bullet \bigoplus_{i \in I} (h_i * \mu_i))(\coprod_{i \in I} A_i) \\
&= \sum_{i \in I} \overline{h}(i)\, (h_i * \mu_i)(A_i) \\
&= \sum_{i \in I} \frac{\int_{X_i} h(i)\, f|_{X_i}\, d\mu_i}{\sum_{i \in I} \int_{X_i} h(i)\, f|_{X_i}\, d\mu_i} \frac{\int_{X_i} \mathbf{1}_{A_i} h_i\, d\mu_i}{\int_{X_i} h_i\, d\mu_i} \\
&= \frac{\sum_{i \in I} \int_{X_i} h(i)\, \mathbf{1}_{A_i}\, f|_{X_i}\, d\mu_i}{\sum_{i \in I} \int_{X_i} h(i)\, f|_{X_i}\, d\mu_i} \\
&= \frac{\int_{\coprod_{i \in I} X_i} \mathbf{1}_{\coprod_{i \in I} A_i} f\, d(h \bullet \bigoplus_{i \in I} \mu_i)}{\int_{\coprod_{i \in I} X_i} f\, d(h \bullet \bigoplus_{i \in I} \mu_i)} \\
&= \frac{\int_{\coprod_{i \in I} X_i} \mathbf{1}_{\coprod_{i \in I} A_i} f\, d\mu}{\int_{\coprod_{i \in I} X_i} f\, d\mu} \\
&= (f * \mu)(\coprod_{i \in I} A_i).
\end{aligned}$$

□

Next are two results concerning sum measures and fusion. The first shows that, given a probability measure $\mu : \mathcal{P}(\coprod_{i \in I} X_i)$ such that $\mu = h \bullet \bigoplus_{i \in I} \mu_i$ and a probability kernel $\eta : \coprod_{i \in I} X_i \to \mathcal{P}(\coprod_{i \in I} X_i)$, there exists a function $\overline{h}$ and, for all $i \in I$, a probability kernel $\eta_i : X_i \to \mathcal{P}(X_i)$ such that $\mu \odot \eta = \overline{h} \bullet \bigoplus_{i \in I} (\mu_i \odot \eta_i)$.

Proposition A.9.14. *Let $(X_i, \mathcal{A}_i)$ be a measurable space, for all $i \in I$, and $\mu : \mathcal{P}(\coprod_{i \in I} X_i)$ a probability measure, where $\mu = h \bullet \bigoplus_{i \in I} \mu_i$, for some $h : I \to [0, \infty)$ and $\mu_i : \mathcal{P}(X_i)$, for all $i \in I$. Let $\eta : \coprod_{i \in I} X_i \to \mathcal{P}(\coprod_{i \in I} X_i)$ be a probability kernel such that $(\mu \odot \eta)(X_i) > 0$, for all $i \in I$, and $\mu \odot \eta$ is absolutely continuous with respect to μ. Then there exists $\overline{h} : I \to [0, \infty)$ and, for all $i \in I$, a probability kernel $\eta_i : X_i \to \mathcal{P}(X_i)$ such that*

$$\mu \odot \eta = \overline{h} \bullet \bigoplus_{i \in I} (\mu_i \odot \eta_i).$$

Proof. By Proposition A.9.9, there exists $\overline{h} : I \to [0, \infty)$ and, for all $i \in I$, a probability measure $\mu_i' : \mathcal{P}(X_i)$ such that $\mu \odot \eta = \overline{h} \bullet \bigoplus_{i \in I} \mu_i'$. Then, by the absolute continuity assumption, it follows that μ_i' is absolutely continuous with respect to μ_i, for all $i \in I$. Thus, by the Radon-Nikodym theorem (Proposition A.2.12), for all $i \in I$, there exists a non-negative integrable $h_i : X_i \to \mathbb{R}$ such that $\mu_i'(B_i) = \int_{X_i} \mathbf{1}_{B_i} h_i\, d\mu_i$, for all $B_i \in \mathcal{A}_i$.

Now define, for all $i \in I$, $\eta_i : X_i \to \mathcal{P}(X_i)$ by $\eta_i(x_i)(B_i) = \mathbf{1}_{B_i}(x_i)h_i(x_i)$, for all $x_i \in X_i$ and $B_i \in \mathcal{A}_i$. The function η_i is well-defined since $\mathbf{1}_{B_i}(x_i)h_i(x_i)$ is a probability measure for fixed x_i. Also η_i is a probability kernel, by Proposition A.2.4. Then $\mu_i'(B_i) = \int_{X_i} \mathbf{1}_{B_i} h_i \, d\mu_i = \int_{X_i} \lambda x_i.\eta_i(x_i)(B_i) \, d\mu_i = (\mu_i \odot \eta_i)(B_i)$, for all $B_i \in \mathcal{A}_i$, and so $\mu_i' = \mu_i \odot \eta_i$.

Finally, $\mu \odot \eta = \overline{h} \bullet \bigoplus_{i\in I} \mu_i' = \overline{h} \bullet \bigoplus_{i\in I}(\mu_i \odot \eta_i)$. □

In a special case, explicit forms for h and each η_i in Proposition A.9.14 can be given.

Proposition A.9.15. *Let $(X_i, \mathcal{A}_i)$ be a measurable space, for all $i \in I$, and $\mu : \mathcal{P}(\coprod_{i\in I} X_i)$ a probability measure, where $\mu = h \bullet \bigoplus_{i\in I} \mu_i$, for some $h : I \to [0, \infty)$ and $\mu_i : \mathcal{P}(X_i)$, for all $i \in I$. Let $\eta : \coprod_{i\in I} X_i \to \mathcal{P}(\coprod_{i\in I} X_i)$ be a probability kernel such that, for all $i \in I$ and $x_i \in X_i$, $\eta(x_i)(X_i) = 1$. For all $i \in I$, define*

$$\eta_i : X_i \to \mathcal{P}(X_i)$$

by

$$\eta_i(x_i)(A_i) = \eta(x_i)(A_i),$$

for all $x_i \in X_i$ and $A_i \in \mathcal{A}_i$. Then

$$\mu \odot \eta = h \bullet \bigoplus_{i\in I}(\mu_i \odot \eta_i).$$

Proof. By Proposition A.9.9, there exists a function $h : I \to [0, \infty)$ such that $\sum_{i\in I} h(i) = 1$ and $\mu_i : \mathcal{P}(X_i)$, for all $i \in I$, such that $\mu = h \bullet \bigoplus_{i\in I} \mu_i$. Also each η_i is well-defined.

For all $\coprod_{i\in I} A_i \in \bigoplus_{i\in I} \mathcal{A}_i$,

$$\begin{aligned}
&(\mu \odot \eta)(\coprod_{i\in I} A_i) \\
&= \int_{\coprod_{i\in I} X_i} \lambda x.\eta(x)(\coprod_{i\in I} A_i) \, d\mu \\
&= \int_{\coprod_{i\in I} X_i} \lambda x.\eta(x)(\coprod_{i\in I} A_i) \, d(h \bullet \bigoplus_{i\in I} \mu_i) \\
&= \sum_{i\in I} \int_{X_i} h(i) \, (\lambda x.\eta(x)(\coprod_{i\in I} A_i))|_{X_i} \, d\mu_i \\
&= \sum_{i\in I} \int_{X_i} h(i) \, \lambda x_i.\eta(x_i)(A_i) \, d\mu_i && [\eta(x_i)(X_i) = 1] \\
&= \sum_{i\in I} \int_{X_i} h(i) \, \lambda x_i.\eta_i(x_i)(A_i) \, d\mu_i \\
&= \sum_{i\in I} h(i) \, (\mu_i \odot \eta_i)(A_i) \\
&= (h \bullet \bigoplus_{i\in I}(\mu_i \odot \eta_i))(\coprod_{i\in I} A_i).
\end{aligned}$$

□

A.10 Quotients of Probability Kernels

Proposition A.10.1. *Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ be measurable spaces, $\mu : X \to \mathcal{P}(Y)$ a probability kernel, and $p : Y \to Z$ a measurable function. Then $\lambda x.(\mu(x) \circ p^{-1}) : X \to \mathcal{P}(Z)$ is a probability kernel.*

Proof. Note first that Proposition A.2.13 shows that $\lambda x.(\mu(x) \circ p^{-1}) : X \to \mathcal{P}(Z)$ is well-defined.

Let $C \in \mathcal{C}$. Then $\lambda x.(\mu(x) \circ p^{-1})(C) = \lambda x.\mu(x)(p^{-1}(C))$ is measurable, by Proposition A.2.4, since μ is a probability kernel and p is measurable. Hence, by Proposition A.2.4 again, $\lambda x.(\mu(x) \circ p^{-1})$ is a probability kernel. □

Definition A.10.1. Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ be measurable spaces, $\mu : X \to \mathcal{P}(Y)$ a probability kernel, and $p : Y \to Z$ a measurable function. Then $\lambda x.(\mu(x) \circ p^{-1}) : X \to \mathcal{P}(Z)$ is called the *quotient* of the probability kernel μ by p, and is denoted by μ/p.

As a special case of Definition A.10.1, if $\mu : \mathcal{P}(Y)$ and $p : Y \to Z$ is a measurable function, then the quotient probability measure $\mu/p : \mathcal{P}(Z)$ is defined by $\mu/p = \mu \circ p^{-1}$.

If the function p in Definition A.10.1 is a surjection, then Z is isomorphic to Y/p, the set of equivalence classes on Y induced by p. Thus (up to isomorphism)

$$\mu/p : X \to \mathcal{P}(Y/p).$$

A useful example of a quotient probability kernel is a marginal probability kernel that is defined in Definition A.2.9.

Now it is shown that the quotient of a regular conditional distribution is also a regular conditional distribution. This result extends Proposition A.7.9.

Proposition A.10.2. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ measurable spaces, $f : \Omega \to X$ and $g : \Omega \to Y$ measurable functions, $p : Y \to Z$ a measurable function, and $\mu : X \to \mathcal{P}(Y)$ a regular conditional distribution of g given f. Then $\mu/p : X \to \mathcal{P}(Z)$ is a regular conditional distribution of $p \circ g$ given f.*

Proof. For all $C \in \mathcal{C}$, almost surely,

$$\begin{aligned}
&\mathsf{P}((p \circ g)^{-1}(C) \mid f) \\
&= \mathsf{P}(g^{-1}(p^{-1}(C)) \mid f) \\
&= \lambda\omega.\mu(f(\omega))(p^{-1}(C)) && [\mu \text{ is a regular conditional distribution}] \\
&= \lambda\omega.\lambda x.(\mu(x) \circ p^{-1})(f(\omega))(C) \\
&= \lambda\omega.(\mu/p)(f(\omega))(C).
\end{aligned}$$

□

Proposition A.10.3. *Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, $(Z, \mathcal{C})$, and $(W, \mathcal{D})$ be measurable spaces, $\mu : X \times Y \to \mathcal{P}(Z)$ a probability kernel, and $p : Z \to W$ a measurable function. Then, for all $x \in X$, $\lambda y.\mu(x, y)/p = \lambda y.(\mu/p)(x, y)$.*

Proof. Note that, for all $x \in X$, by Proposition A.2.5, $\lambda y.\mu(x,y) : Y \to \mathcal{P}(Z)$ is a probability kernel, so that, by Proposition A.10.1, $\lambda y.\mu(x,y)/p : Y \to \mathcal{P}(W)$ is also a probability kernel.

Now, for all $x \in X$ and $D \in \mathcal{D}$,

$$\begin{aligned}
&(\lambda y.\mu(x,y)/p)(y)(D) \\
&= \lambda y.(\mu(x,y) \circ p^{-1})(y)(D) \\
&= \lambda y.(\lambda(x,y).(\mu(x,y) \circ p^{-1})(x,y))(y)(D) \\
&= \lambda y.(\mu/p)(x,y)(y)(D).
\end{aligned}$$

□

Proposition A.10.4. *Let $(X,\mathcal{A})$, $(Y,\mathcal{B})$, and $(Z,\mathcal{C})$ be measurable spaces, $\mu : X \to \mathcal{P}(Y)$ a probability kernel, and $p : Y \to Z$ a measurable function. Suppose that, for all $x \in X$, $g : Z \to \mathbb{R}$ is $(\mu/p)(x)$-integrable. Then, for all $x \in X$,*

$$\int_Z g \; d(\mu/p)(x) = \int_Y (g \circ p) \; d\mu(x).$$

Proof. Since $(\mu/p)(x) = \mu(x) \circ p^{-1}$, the result follows immediately from Proposition A.2.14. □

As promised above, the use of one point extensions in Section A.9 is now reinterpreted in the light of quotient kernels. Consider a probability kernel $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$. For all $i \in I$, the equivalence relation $\sim_i$ on $\coprod_{i\in I} X_i$ is defined by

$$x \sim_i y = \begin{cases} T & \text{if } x, y \notin X_i \\ F & \text{otherwise,} \end{cases}$$

for all $x, y \in \coprod_{i\in I} X_i$. Then, for all $i \in I$, $\coprod_{i\in I} X_i/\sim_i$ denotes the set of equivalence classes under $\sim_i$ and $[x]_i$ denotes the equivalence class containing x. Clearly, each $\coprod_{i\in I} X_i/\sim_i$ is isomorphic to $X_i \sqcup \{*\}$, where $\{*\}$ corresponds to the equivalence class $\coprod_{j\in I,\, j\neq i} X_j$.

Now, for all $i \in I$, define $p_i : \coprod_{i\in I} X_i \to \coprod_{i\in I} X_i/\sim_i$ by $p_i(x) = [x]_i$, for all $x \in \coprod_{i\in I} X_i$. Then each p_i is a measurable surjection. Identifying each $\coprod_{i\in I} X_i/\sim_i$ and $X_i \sqcup \{*\}$, respectively, the quotient probability kernel $\mu/p_i : X \to \mathcal{P}(\coprod_{i\in I} X_i/\sim_i)$ is the same as the probability kernel $\mu_i : X \to \mathcal{P}(X_i \sqcup \{*\})$ in Proposition A.9.2. Thus $\mu = \bigoplus_{i\in I} \mu/p_i$, and so μ is a sum of quotient probability kernels.

Next is an important application of quotient kernels. Consider a probability kernel μ having signature of the form $X \to \mathcal{P}(W^Z)$, where X and W are measurable sets and Z is a set. If Z is an infinite set, there is a practical problem in dealing with distributions on the infinite product space W^Z. The next result, together with a further finiteness condition, shows how to construct probability kernels that are practicable for a large class of typical applications.

Proposition A.10.5. *Let $(\Omega,\mathfrak{S},\mathsf{P})$ be a probability space, $(X,\mathcal{A})$ and $(W,\mathcal{D})$ measurable spaces, Z a set, $\sim$ an equivalence relation on Z, $f : \Omega \to X$ and $g : \Omega \to W^{Z/\sim}$ measurable*

functions, $\mu : X \to \mathcal{P}(W^{Z/\sim})$ a probability kernel, and $\pi : Z \to Z/\sim$ the canonical surjection. Define $p : W^{Z/\sim} \to W^Z$ by

$$p(h) = h \circ \pi,$$

for all $h \in W^{Z/\sim}$. Let C be the set of functions in W^Z that are constant on each equivalence class in the partition of Z. Then the following hold.

1. *p is measurable.*
2. *p is a bijection between $W^{Z/\sim}$ and C.*
3. *$\lambda x.(\mu(x) \circ p^{-1}) : X \to \mathcal{P}(W^Z)$ is a probability kernel.*
4. *If μ is a regular conditional distribution of g given f, then $\lambda x.(\mu(x) \circ p^{-1})$ is a regular conditional distribution of $p \circ g$ given f.*
5. *For all $x \in X$, $(\mu(x) \circ p^{-1})(C) = 1$.*
6. *For all $x \in X$, if $k : W^Z \to \mathbb{R}$ is a $(\mu(x) \circ p^{-1})$-integrable function, then*

$$\int_{W^Z} k \, d(\mu(x) \circ p^{-1}) = \int_{W^{Z/\sim}} k \circ p \, d\mu(x).$$

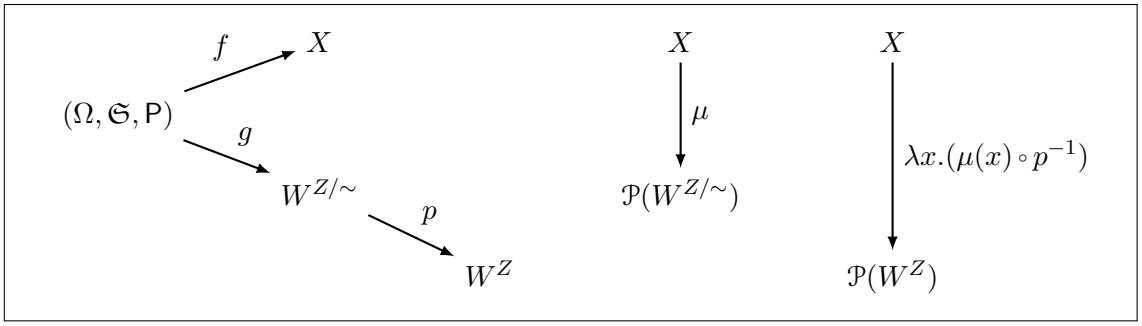

Figure A.40: Setting for Proposition A.10.5

Proof. 1. By Proposition A.1.5, to show that p is measurable, it suffices to show that $\pi_z \circ p : W^{Z/\sim} \to W$ is measurable, for each evaluation map $\pi_z : W^Z \to W$. For this, it suffices to show that $\pi_z \circ p = \pi_{[z]}$, for all $z \in Z$. (Here $[z]$ is the equivalence class containing z and $\pi_{[z]} : W^{Z/\sim} \to W$ is the evaluation map at $[z]$.) Now, for all $h \in W^{Z/\sim}$, $(\pi_z \circ p)(h) = \pi_z(p(h)) = \pi_z(h \circ \pi) = h(\pi(z)) = h([z])$. Thus $\pi_z \circ p = \pi_{[z]}$, for all $z \in Z$.

2. Suppose that $h_1, h_2 \in W^{Z/\sim}$, where $h_1 \neq h_2$. Hence there exists $z \in Z$ such that $h_1(\pi(z)) \neq h_2(\pi(z))$, and so $p(h_1) \neq p(h_2)$. Thus p is injective.

Suppose that $k \in C$. Thus k is constant on each equivalence in the partition of Z. Consequently, it is possible to factor k so that $k = h \circ \pi$, for some $h \in W^{Z/\sim}$. Thus $p(h) = k$ and so p is surjective.

3. This part follows from Part 1 and Proposition A.10.1.

4. This part follows immediately from Proposition A.10.2.

5. For all $x \in X$, $(\mu(x) \circ p^{-1})(C) = \mu(x)(p^{-1}(C)) = \mu(x)(W^{Z/\sim}) = 1$.

6. This part follows immediately from Proposition A.2.14. □

Proposition A.10.5 provides a method for obtaining a useful class of probability kernels having signature of the form $X \to \mathcal{P}(W^Z)$. This is especially important when X is the set of histories up to the current time and W^Z is being thought of a function space, say, a space of hypotheses for learning a function having signature $Z \to W$, rather than as a product space. Suppose that Z is infinite. Then define a suitable *finite* partition on Z (using, for example, predicates generated by a predicate rewrite system of Section B.1.12). This means that $Z/\!\sim$ is finite and hence it is much easier to construct probability kernels having signature $X \to \mathcal{P}(W^{Z/\sim})$. Now use Proposition A.10.5 to lift such probability kernels to probability kernels having signature $X \to \mathcal{P}(W^Z)$. Then, for each $x \in X$, one has a distribution over the space of hypotheses. Note that Part 5 of Proposition A.10.5 shows that the support of such distributions is a set of piecewise-constant functions in W^Z. This is fine for many learning settings, especially when Z is structured. (Of course, everything works the same if one thinks of W^Z as an infinite product space.) If $Z/\!\sim$ is finite, then Part 6 shows that it is possible to compute integrals of real-valued functions on W^Z with respect to the probability measure $\mu(x) \circ p^{-1}$, for all $x \in X$, since $W^{Z/\sim}$ is a finite product space. This can be done efficiently by Monte Carlo methods.

Here is another way of constructing distributions on a function space W^Z, this time exploiting a linearity condition. Note that $\mathbb{R}^{\mathbb{R}^m}$ means $\mathbb{R}^{(\mathbb{R}^m)}$ (not $(\mathbb{R}^{\mathbb{R}})^m$).

Proposition A.10.6. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X, \mathcal{A})$ a measurable space, $f : \Omega \to X$ and $g : \Omega \to \mathbb{R}^{m+1}$, for some $m \in \mathbb{N}$, measurable functions, and $\mu : X \to \mathcal{P}(\mathbb{R}^{m+1})$ a probability kernel. Define $p : \mathbb{R}^{m+1} \to \mathbb{R}^{\mathbb{R}^m}$ by*

$$p(a_1, \ldots, a_{m+1}) = \lambda(x_1, \ldots, x_m).(\sum_{j=1}^{m} a_j x_j + a_{m+1}),$$

for all $(a_1, \ldots, a_{m+1}) \in \mathbb{R}^{m+1}$. Let

$$L = \{f \in \mathbb{R}^{\mathbb{R}^m} \mid f = \lambda(x_1, \ldots, x_m).(\sum_{j=1}^{m} a_j x_j + a_{m+1}), \text{ for some } (a_1, \ldots, a_{m+1}) \in \mathbb{R}^{m+1}\}.$$

Then the following hold.

1. *p is measurable.*
2. *$\lambda x.(\mu(x) \circ p^{-1}) : X \to \mathcal{P}(\mathbb{R}^{\mathbb{R}^m})$ is a probability kernel.*
3. *If μ is a regular conditional distribution of g given f, then $\lambda x.(\mu(x) \circ p^{-1})$ is a regular conditional distribution of $p \circ g$ given f.*
4. *For all $x \in X$, $(\mu(x) \circ p^{-1})(L) = 1$.*
5. *For all $x \in X$, if $f : \mathbb{R}^{\mathbb{R}^m} \to \mathbb{R}$ is a $(\mu(x) \circ p^{-1})$-integrable function, then*

$$\int_{\mathbb{R}^{\mathbb{R}^m}} f \, d(\mu(x) \circ p^{-1}) = \int_{\mathbb{R}^{m+1}} f \circ p \, d\mu(x).$$

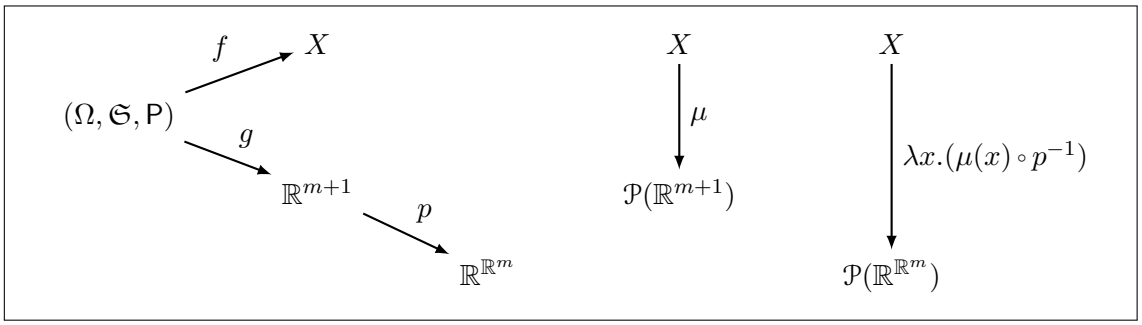

Figure A.41: Setting for Proposition A.10.6

Proof. 1. By Proposition A.1.5, to show that p is measurable, it suffices to show that $\pi_x \circ p : \mathbb{R}^{m+1} \to \mathbb{R}$ is measurable, for each evaluation map $\pi_x : \mathbb{R}^{\mathbb{R}^m} \to \mathbb{R}$, where $x \triangleq (x_1, \ldots, x_m) \in \mathbb{R}^m$. Now $(\pi_x \circ p)(a_1, \ldots, a_{m+1}) = \sum_{j=1}^{m} a_j x_j + a_{m+1}$, for all $(a_1, \ldots, a_{m+1}) \in \mathbb{R}^{m+1}$. Thus $\pi_x \circ p$ is continuous and hence measurable.

2. This part follows from Part 1 and Proposition A.10.1.
3. This part follows immediately from Proposition A.10.2.
4. For all $x \in X$, $(\mu(x) \circ p^{-1})(L) = \mu(x)(p^{-1}(L)) = \mu(x)(\mathbb{R}^{m+1}) = 1$.
5. This part follows immediately from Proposition A.2.14. □

Now a connection between quotient probability measures and projective products is investigated.

Proposition A.10.7. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $p : X \to Y$ a measurable function, $\mu : \mathcal{P}(X)$ a probability measure, and $f : X \to \mathbb{R}$ a non-negative measurable function such that $0 < \int_X f \, d\mu < \infty$. Suppose that there exists a measurable function $\overline{f} : Y \to \mathbb{R}$ such that $f = \overline{f} \circ p$. Then*

$$(f * \mu)/p = \overline{f} * (\mu/p).$$

Proof. Note that $\int_Y \overline{f} \, d(\mu/p) = \int_X f \, d\mu$ and hence $0 < \int_Y \overline{f} \, d(\mu/p) < \infty$. Thus $\overline{f} * (\mu/p)$ is well-defined.

For all $B \in \mathcal{B}$,

$$\begin{aligned}
&((f * \mu)/p)(B) \\
&= ((f * \mu) \circ p^{-1})(B) \\
&= \frac{\int_X \mathbf{1}_{p^{-1}(B)} f \, d\mu}{\int_X f \, d\mu} \\
&= \frac{\int_X \mathbf{1}_{p^{-1}(B)} (\overline{f} \circ p) \, d\mu}{\int_X (\overline{f} \circ p) \, d\mu} \\
&= \frac{\int_Y \mathbf{1}_B \overline{f} \, d(\mu \circ p^{-1})}{\int_Y \overline{f} \, d(\mu \circ p^{-1})} \\
&= \frac{\int_Y \mathbf{1}_B \overline{f} \, d(\mu/p)}{\int_Y \overline{f} \, d(\mu/p)} \\
&= (\overline{f} * (\mu/p))(B).
\end{aligned}$$

□

Next quotient probability measures and fusion are studied.

Proposition A.10.8. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $p : X \to Y$ a measurable surjection, and $\mu : \mathcal{P}(X)$ a probability measure. Let $\eta : X \to \mathcal{P}(X)$ be a probability kernel such that $\eta = \overline{\eta} \circ p$, for some probability kernel $\overline{\eta} : Y \to \mathcal{P}(X)$. Then*

$$(\mu \odot \eta)/p = (\mu/p) \odot (\overline{\eta}/p).$$

Proof. For all $B \in \mathcal{B}$,

$$\begin{aligned}
&((\mu \odot \eta)/p)(B) \\
&= ((\mu \odot \eta) \circ p^{-1})(B) \\
&= \int_X \lambda x.\eta(x)(p^{-1}(B))\, d\mu \\
&= \int_X \lambda x.\overline{\eta}(p(x))(p^{-1}(B))\, d\mu \\
&= \int_X \lambda x.(\overline{\eta}/p)(p(x))(B)\, d\mu \\
&= \int_X (\lambda y.(\overline{\eta}/p)(y)(B) \circ p)\, d\mu \\
&= \int_Y \lambda y.(\overline{\eta}/p)(y)(B)\, d(\mu \circ p^{-1}) \\
&= ((\mu/p) \odot (\overline{\eta}/p))(B).
\end{aligned}$$

□

A.11 Restrictions of Probability Kernels

Proposition A.11.1. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $\mu : X \to \mathcal{P}(Y)$ a probability kernel, and $Z \in \mathcal{B}$ such that $\mu(x)(Z) > 0$, for all $x \in X$. Define $\mu||_Z : X \to \mathcal{P}(Z)$ by*

$$\mu||_Z(x) = \mu(x)|_Z,$$

for all $x \in X$. Then $\mu||_Z : X \to \mathcal{P}(Z)$ is a probability kernel.

Proof. More explicitly, $\mu||_Z$ is defined by

$$\mu||_Z(x)(B \cap Z) = \frac{\mu(x)(B \cap Z)}{\mu(x)(Z)},$$

for all $B \in \mathcal{B}$ and $x \in X$. By Proposition A.2.4, to show that $\mu||_Z$ is a probability kernel it suffices to show that

$$\lambda x.\frac{\mu(x)(B \cap Z)}{\mu(x)(Z)} : X \to \mathbb{R},$$

is measurable, for all $B \in \mathcal{B}$. But this follows directly from the fact that μ is a probability kernel. □

Definition A.11.1. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $\mu : X \to \mathcal{P}(Y)$ a probability kernel, and $Z \in \mathcal{B}$ such that $\mu(x)(Z) > 0$, for all $x \in X$. Then $\mu||_Z : X \to \mathcal{P}(Z)$ is called the *restriction* of μ to Z.

The next result is used for computing integrals with respect to restrictions of probability kernels.

Proposition A.11.2. *Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $\mu : X \to \mathcal{P}(Y)$ a probability kernel, $Z \in \mathcal{B}$ such that $\mu(x)(Z) > 0$, for all $x \in X$, and $\mu||_Z : X \to \mathcal{P}(Z)$ the restriction of μ to Z. Suppose that $f : Z \to \mathbb{R}$ is a non-negative measurable function. Then, for all $x \in X$,*

$$\int_Z f \; d\mu||_Z(x) = \frac{\int_Y \bar{f} \; d\mu(x)}{\mu(x)(Z)},$$

where $\bar{f}$ is f extended to Y by defining it to be 0 on $Y \setminus Z$.

Proof. The result holds when f is a measurable indicator function, and hence holds for simple functions. Now apply the monotone convergence theorem (Proposition A.2.2). □

A.12 Products of Conditional Densities

First, products of conditional densities are defined.

Definition A.12.1. Let $(X_0, \mathcal{A}_0)$ be a measurable space, $(X_1, \mathcal{A}_1, \nu_1)$ and $(X_2, \mathcal{A}_2, \nu_2)$ measure spaces, and $h_1 : X_0 \to \mathcal{D}(X_1)$ and $h_2 : X_0 \times X_1 \to \mathcal{D}(X_2)$ conditional densities. Then the *product* $h_1 \otimes h_2 : X_0 \to \mathcal{D}(X_1 \times X_2)$ of h_1 and h_2 is defined by

$$(h_1 \otimes h_2)(x_0) = \lambda(x_1, x_2).h_1(x_0)(x_1)\, h_2(x_0, x_1)(x_2),$$

for all $x_0 \in X_0$.

The product is indeed a conditional density.

Proposition A.12.1. *Let $(X_0, \mathcal{A}_2)$ be a measurable space, $(X_1, \mathcal{A}_1, \nu_1)$, and $(X_2, \mathcal{A}_2, \nu_2)$ σ-finite measure spaces, and $h_1 : X_0 \to \mathcal{D}(X_1)$ and $h_2 : X_0 \times X_1 \to \mathcal{D}(X_2)$ conditional densities. Then $h_1 \otimes h_2 : X_0 \to \mathcal{D}(X_1 \times X_2)$ is a conditional density.*

Proof. The function $h_1 \otimes h_2$ is well-defined, since, for all $x_0 \in X_0$, $(h_1 \otimes h_2)(x_0)$ is measurable and

$$\begin{aligned}
&\int_{X_1 \times X_2} (h_1 \otimes h_2)(x_0) \; d(\nu_1 \otimes \nu_2) \\
&= \int_{X_1 \times X_2} \lambda(x_1, x_2).h_1(x_0)(x_1)\, h_2(x_0, x_1)(x_2) \; d(\nu_1 \otimes \nu_2) \\
&= \int_{X_1} \lambda x_1.h_1(x_0)(x_1) \left(\int_{X_2} \lambda x_2.h_2(x_0, x_1)(x_2) \; d\nu_2 \right) \; d\nu_1 && \text{[Proposition A.2.16]} \\
&= \int_{X_1} \lambda x_1.h_1(x_0)(x_1) \; d\nu_1 && [h_2 \text{ is a cond. density}] \\
&= 1. && [h_1 \text{ is a cond. density}]
\end{aligned}$$

Also the mapping $(x_0, x_1, x_2) \mapsto h_1(x_0)(x_1)\, h_2(x_0, x_1)(x_2)$ is measurable, since h_1 and h_2 are conditional densities. Thus $h_1 \otimes h_2$ is a conditional density. □

Note. Products and fusions of probability densities are closely related:

$$(h_1 \odot h_2)(x_0) = \lambda x_2. \int_{X_1} \lambda x_1.(h_1 \otimes h_2)(x_0)(x_1, x_2)\, d\nu_1,$$

for all $x_0 \in X_0$.

Suppose that

$$\begin{aligned}
&h_1 : X_0 \to \mathcal{D}(X_1)\\
&h_2 : X_0 \times X_1 \to \mathcal{D}(X_2)\\
&h_3 : X_0 \times X_1 \times X_2 \to \mathcal{D}(X_3)\\
&\quad\vdots\\
&h_{n-1} : X_0 \times X_1 \times \cdots \times X_{n-2} \to \mathcal{D}(X_{n-1})\\
&h_n : X_0 \times X_1 \times \cdots \times X_{n-2} \times X_{n-1} \to \mathcal{D}(X_n)
\end{aligned}$$

are conditional densities. Then $\bigotimes_{i=1}^n h_i$ means $(\cdots(((h_1 \otimes h_2) \otimes h_3) \otimes h_4) \cdots \otimes h_n)$, and is well-defined, by Proposition A.12.1.

Proposition A.12.2. *Let $(X_0, \mathcal{A}_0)$ be a measurable space, $(X_i, \mathcal{A}_i, \nu_i)$ a σ-finite measure space, for $i = 1, \ldots, n$, and $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ a conditional density, for $i = 1, \ldots, n$. Then $\bigotimes_{i=1}^n h_i : X_0 \to \mathcal{D}(\prod_{i=1}^n X_i)$ is a conditional density.*

Proof. By induction on n, using Proposition A.12.1. □

The next result gives an simple connection between conditional densities, measures, and products.

Proposition A.12.3. *Let $(X_0, \mathcal{A}_0)$ be a measurable space, $(X_i, \mathcal{A}_i, \nu_i)$ a σ-finite measure space, for $i = 1, \ldots, n$, and $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ a conditional density, for $i = 1, \ldots, n$. Then*

$$\bigotimes_{i=1}^n (h_i \cdot \nu_i) = (\bigotimes_{i=1}^n h_i) \cdot (\bigotimes_{i=1}^n \nu_i).$$

Proof. Let $\mathcal{P} \triangleq \{\prod_{i=1}^n A_i \mid A_i \in \mathcal{A}_i, \text{ for } i = 1, \ldots, n\}$ and

$$\mathcal{L} \triangleq \{A \in \bigotimes_{i=1}^n \mathcal{A}_i \mid (\bigotimes_{i=1}^n (h_i \cdot \nu_i))(x_0)(A) = ((\bigotimes_{i=1}^n h_i) \cdot (\bigotimes_{i=1}^n \nu_i))(x_0)(A), \text{ for all } x_0 \in X_0\}.$$

Note that $\mathcal{P}$ is a π-system and $\sigma(\mathcal{P}) = \bigotimes_{i=1}^n \mathcal{B}_i$.

Suppose that $\prod_{i=1}^n A_i \in \mathcal{P}$. It is shown by induction on n that $\prod_{i=1}^n A_i \in \mathcal{L}$. For $n = 1$, the result is obvious. For the inductive step, let $x_0 \in X_0$ and $A_i \in \mathcal{A}_i$, for $i = 1, \ldots, n$.

Then

$$
\begin{aligned}
&(\bigotimes_{i=1}^{n}(h_i \cdot \nu_i))(x_0)(\prod_{i=1}^{n} A_i) \\
&= ((\bigotimes_{i=1}^{n-1}(h_i \cdot \nu_i)) \otimes (h_n \cdot \nu_n))(x_0)(\prod_{i=1}^{n} A_i) \\
&= ((((\bigotimes_{i=1}^{n-1} h_i) \cdot (\bigotimes_{i=1}^{n-1} \nu_i)) \otimes (h_n \cdot \nu_n))(x_0)(\prod_{i=1}^{n} A_i) \quad \text{[Induction hypothesis]} \\
&= \int_{\prod_{i=1}^{n-1} X_i} \mathbf{1}_{\prod_{i=1}^{n-1} A_i} \lambda(x_1, \ldots, x_{n-1}).(h_n \cdot \nu_n)(x_0, \ldots, x_{n-1})(A_n)\; d((\bigotimes_{i=1}^{n-1} h_i) \cdot (\bigotimes_{i=1}^{n-1} \nu_i))(x_0) \\
&= \int_{\prod_{i=1}^{n-1} X_i} \mathbf{1}_{\prod_{i=1}^{n-1} A_i} (\lambda(x_1, \ldots, x_{n-1}).(h_n \cdot \nu_n)(x_0, \ldots, x_{n-1})(A_n))(\bigotimes_{i=1}^{n-1} h_i)(x_0)\; d\bigotimes_{i=1}^{n-1} \nu_i \\
&= \int_{\prod_{i=1}^{n-1} X_i} \mathbf{1}_{\prod_{i=1}^{n-1} A_i} \left(\lambda(x_1, \ldots, x_{n-1}). \int_{X_n} \mathbf{1}_{A_n} h_n(x_0, \ldots, x_{n-1})\; d\nu_n\right) (\bigotimes_{i=1}^{n-1} h_i)(x_0)\; d\bigotimes_{i=1}^{n-1} \nu_i \\
&= \int_{\prod_{i=1}^{n-1} X_i} \left(\lambda(x_1, \ldots, x_{n-1}). \int_{X_n} \lambda x_n.(\mathbf{1}_{\prod_{i=1}^{n} A_i}(x_1, \ldots, x_n) \right. \\
&\qquad \left. (\bigotimes_{i=1}^{n-1} h_i)(x_0)(x_1, \ldots, x_{n-1})\, h_n(x_0, \ldots, x_{n-1})(x_n))\; d\nu_n\right)\; d\bigotimes_{i=1}^{n-1} \nu_i \\
&= \int_{\prod_{i=1}^{n-1} X_i} \left(\lambda(x_1, \ldots, x_{n-1}). \int_{X_n} \lambda x_n.(\mathbf{1}_{\prod_{i=1}^{n} A_i} (\bigotimes_{i=1}^{n} h_i)(x_0))(x_1, \ldots, x_n)\; d\nu_n\right)\; d\bigotimes_{i=1}^{n-1} \nu_i \\
&= \int_{\prod_{i=1}^{n} X_i} \mathbf{1}_{\prod_{i=1}^{n} A_i} (\bigotimes_{i=1}^{n} h_i)(x_0)\; d\bigotimes_{i=1}^{n} \nu_i \quad \text{[Proposition A.2.16]} \\
&= ((\bigotimes_{i=1}^{n} h_i) \cdot (\bigotimes_{i=1}^{n} \nu_i))(x_0)(\prod_{i=1}^{n} A_i).
\end{aligned}
$$

Thus $\prod_{i=1}^{n} A_i \in \mathcal{L}$ and so $\mathcal{P} \subseteq \mathcal{L}$.

Next it is shown that $\mathcal{L}$ is a λ-system. First, $\prod_{i=1}^{n} X_i \in \mathcal{P}$ and $\mathcal{P} \subseteq \mathcal{L}$, so that $\prod_{i=1}^{n} X_i \in \mathcal{L}$.

Second, let $A, B \in \mathcal{L}$, where $A \subseteq B$. Then, for all $x_0 \in X_0$,

$$
\begin{aligned}
&\bigotimes_{i=1}^{n}(h_i \cdot \nu_i)(x_0)(B \setminus A) \\
&= \bigotimes_{i=1}^{n}(h_i \cdot \nu_i)(x_0)(B) - \bigotimes_{i=1}^{n}(h_i \cdot \nu_i)(x_0)(A) \\
&= ((\bigotimes_{i=1}^{n} h_i) \cdot (\bigotimes_{i=1}^{n} \nu_i))(x_0)(B) - ((\bigotimes_{i=1}^{n} h_i) \cdot (\bigotimes_{i=1}^{n} \nu_i))(x_0)(A) \\
&= ((\bigotimes_{i=1}^{n} h_i) \cdot (\bigotimes_{i=1}^{n} \nu_i))(x_0)(B \setminus A).
\end{aligned}
$$

Hence $B \setminus A \in \mathcal{L}$.

Third, let $(B_k)_{k\in\mathbb{N}}$ be an increasing sequence of sets in $\mathcal{L}$. Then, for all $x_0 \in X_0$,

$$\begin{aligned}
&\bigotimes_{i=1}^{n}(h_i \cdot \nu_i)(x_0)(\bigcup_{k\in\mathbb{N}} B_k) \\
&= \lim_{k\to\infty} \bigotimes_{i=1}^{n}(h_i \cdot \nu_i)(x_0)(B_k) \\
&= \lim_{k\to\infty} \bigotimes_{i=1}^{n}(h_i \cdot \nu_i)(x_0)(B_k) \\
&= (\bigotimes_{i=1}^{n} h_i) \cdot (\bigotimes_{i=1}^{n} \nu_i)(x_0)(\bigcup_{k\in\mathbb{N}} B_k)
\end{aligned}$$

Hence $\bigcup_{k\in\mathbb{N}} B_k \in \mathcal{L}$.

It now follows from the monotone-class theorem (Proposition A.1.2) that $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. Hence the result. □

Proposition A.12.4. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, and $(X_i, \mathcal{A}_i, \nu_i)$ a σ-finite measure space, for $i = 1, \ldots, n$. Suppose that $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{i=1}^{n} X_i$ are measurable, and, for $i = 1, \ldots n$, $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ is a conditional density such that*

$$\mathsf{P}(g_i^{-1}(A_i) \mid (f_0, \ldots, f_{i-1})) = \lambda\omega.(h_i \cdot \nu_i)((f_0, \ldots, f_{i-1})(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$. Then

$$\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.((\bigotimes_{i=1}^{n} h_i) \cdot (\bigotimes_{i=1}^{n} \nu_i))(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^{n} \mathcal{A}_i$.

Proof. Let $A \in \bigotimes_{i=1}^{n} \mathcal{A}_i$. Then, almost surely,

$$\begin{aligned}
&\mathsf{P}(f^{-1}(A) \mid f_0) \\
&= \lambda\omega.(\bigotimes_{i=1}^{n}(h_i \cdot \nu_i))(f_0(\omega))(A) && \text{[Proposition A.7.12]} \\
&= \lambda\omega.((\bigotimes_{i=1}^{n} h_i) \cdot (\bigotimes_{i=1}^{n} \nu_i))(f_0(\omega))(A). && \text{[Proposition A.12.3]}
\end{aligned}$$

□

Now the problem of factoring a conditional density mapping into a space of densities on a product space is considered.

Proposition A.12.5. *Let $(X_0, \mathcal{A}_0)$ be a measurable space, $(X_i, \mathcal{A}_i, \nu_i)$ a σ-finite measure space, for $i = 1, \ldots, n$, and $h : X_0 \to \mathcal{D}(\prod_{i=1}^n X_i)$ a conditional density. For $i = 1, \ldots, n$, define $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ by*

$$h_i(x_0, \ldots, x_{i-1}) = \frac{\lambda x_i . \int_{\prod_{j=i+1}^n X_j} \lambda(x_{i+1}, \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=i+1}^n \nu_j}{\int_{\prod_{j=i}^n X_j} \lambda(x_i, \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=i}^n \nu_j},$$

for all $(x_0, \ldots, x_{i-1}) \in \prod_{j=0}^{i-1} X_j$ (where each denominator is assumed to never be 0). Then the following hold.

1. *For $i = 1, \ldots, n$, h_i is conditional density.*
2. *$h = \bigotimes_{i=1}^n h_i$.*

Proof. 1. For $i = 1, \ldots, n$, $j = 0, \ldots, i-1$, and all $x_j \in X_j$, $h_i(x_0, \ldots, x_{i-1})$ is measurable being a quotient of a numerator that is a measurable function, by Proposition A.2.15, and a denominator that is a constant. Also

$$\begin{aligned}
&\int_{X_i} h_i(x_0, \ldots, x_{i-1})\, d\nu_i \\
&= \int_{X_i} \frac{\lambda x_i . \int_{\prod_{j=i+1}^n X_j} \lambda(x_{i+1}, \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=i+1}^n \nu_j}{\int_{\prod_{j=i}^n X_j} \lambda(x_i, \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=i}^n \nu_j}\, d\nu_i \\
&= 1.
\end{aligned}$$

Thus $h_i(x_0, \ldots, x_{i-1})$ is a density.

Furthermore, for $i = 1, \ldots, n$, h_i is measurable since, for all $x_i \in X_i$,

$$\begin{aligned}
&\lambda(x_0, \ldots, x_{i-1}).h_i(x_0, \ldots, x_{i-1})(x_i) \\
&= \frac{\lambda(x_0, \ldots, x_i). \int_{\prod_{j=i+1}^n X_j} \lambda(x_{i+1}, \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=i+1}^n \nu_j}{\lambda(x_0, \ldots, x_{i-1}). \int_{\prod_{j=i}^n X_j} \lambda(x_i, \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=i}^n \nu_j}
\end{aligned}$$

is measurable being a quotient of measurable functions, by Proposition A.2.15; hence h_i is a conditional density.

2. For $i = 0, \ldots, n$, and all $x_i \in X_i$,

$$\begin{aligned}
&(\bigotimes_{i=1}^n h_i)(x_0)(x_1, \ldots, x_n) \\
&= h_1(x_0)(x_1)\, h_2(x_0, x_1)(x_2) \ldots h_n(x_0, \ldots, x_{n-1})(x_n) \\
&= \frac{\int_{\prod_{j=2}^n X_j} \lambda(x_2, \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=2}^n \nu_j}{1} \times \\
&\quad \frac{\int_{\prod_{j=3}^n X_j} \lambda(x_3), \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=3}^n \nu_j}{\int_{\prod_{j=2}^n X_j} \lambda(x_2, \ldots, x_n).h(x_0)(x_1, \ldots, x_n)\, d \bigotimes_{j=2}^n \nu_j} \times \\
&\qquad\qquad \vdots
\end{aligned}$$

$$\begin{aligned}&\frac{h(x_0)(x_1,\ldots,x_n)}{\int_{X_n}\lambda x_n.h(x_0)(x_1,\ldots,x_n)\;d\nu_n}\times\\&h(x_0)(x_1,\ldots,x_n)\\=\;&h(x_0)(x_1,\ldots,x_n).\end{aligned}$$

Hence $h=\bigotimes_{i=1}^{n}h_i$. □

Example A.12.1. For the case $n=2$, Proposition A.12.5 becomes:

$$\begin{aligned}&h:X_0\to\mathcal{D}(X_1\times X_2)\\&h_1:X_0\to\mathcal{D}(X_1)\\&h_1(x_0)=\lambda x_1.\int_{X_2}\lambda x_2.h(x_0)(x_1,x_2)\;d\nu_2,\quad\text{for all } x_0\in X_0\\&h_2:X_0\times X_1\to\mathcal{D}(X_2)\\&h_2(x_0,x_1)=\frac{\lambda x_2.h(x_0)(x_1,x_2)}{\int_{X_2}\lambda x_2.h(x_0)(x_1,x_2)\;d\nu_2},\quad\text{for all } x_0\in X_0 \text{ and } x_1\in X_1\\&h=h_1\otimes h_2.\end{aligned}$$

Taking X_0 to be a singleton set,

$$h_2(x_1)(x_2)=\frac{h(x_1,x_2)}{\int_{X_2}\lambda x_2.h(x_1,x_2)\;d\nu_2},\quad\text{for all } x_1\in X_1 \text{ and } x_2\in X_2,$$

which recovers the usual definition of a conditional probability density function [163, Definition 2.36].

Proposition A.12.6. *Let $(\Omega,\mathfrak{S},\mathsf{P})$ be a probability space, $(X_0,\mathcal{A}_0)$ be a measurable space, $(X_i,\mathcal{A}_i,\nu_i)$ a σ-finite measure space, for $i=1,\ldots,n$, $f_i:\Omega\to X_i$ a random variable, for $i=0,\ldots,n$, and $h_i:\prod_{j=0}^{i-1}X_i\to\mathcal{D}(X_i)$ a conditional density, for $i=1,\ldots,n$. Then $\bigotimes_{i=1}^{n}h_i$ is a regular conditional density with respect to $\bigotimes_{i=1}^{n}\nu_i$ of $(f_1,\ldots,f_n)$ given f_0 iff h_i is a regular conditional density with respect to ν_i of f_i given $(f_0,\ldots,f_{i-1})$, for $i=1,\ldots,n$.*

Proof.

$\bigotimes_{i=1}^{n}h_i$ is a regular conditional density with respect to $\bigotimes_{i=1}^{n}\nu_i$ of $(f_1,\ldots,f_n)$ given f_0

iff $\bigotimes_{i=1}^{n}h_i\cdot\bigotimes_{i=1}^{n}\nu_i$ is a regular conditional distribution of $(f_1,\ldots,f_n)$ given f_0

[Definition A.5.13]

iff $\bigotimes_{i=1}^{n}(h_i\cdot\nu_i)$ is a regular conditional distribution of $(f_1,\ldots,f_n)$ given f_0

[Proposition A.12.3]

iff $h_i\cdot\nu_i$ is a regular conditional distribution of f_i given $(f_0,\ldots,f_{i-1})$, for $i=1,\ldots,n$

[Propositions A.7.12 and A.7.16]

iff h_i is a regular conditional density with respect to ν_i of f_i given $(f_0, \ldots, f_{i-1})$, for $i = 1, \ldots, n$.

[Definition A.5.13]

□

Here is the third, and final, form of Bayes theorem, this time for conditional densities.

Proposition A.12.7. *(Bayes theorem for conditional densities) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, $(X_1, \mathcal{A}_1, \nu_1)$ and $(X_2, \mathcal{A}_2, \nu_2)$ σ-finite measure spaces, and $f_0 : \Omega \to X_0$, $f_1 : \Omega \to X_1$, and $f_2 : \Omega \to X_2$ random variables. Suppose that $h_1 : X_0 \to \mathcal{D}(X_1)$ is a regular conditional density (with respect to ν_1) of f_1 given f_0 and $h_{1,2} : X_0 \times X_1 \to \mathcal{D}(X_2)$ is a regular conditional density (with respect to ν_2) of f_2 given (f_0, f_1). Then there exist a regular conditional density (with respect to ν_2) $h_2 : X_0 \to \mathcal{D}(X_2)$ of f_2 given f_0 defined by*

$$h_2(x_0) = \lambda x_2. \int_{X_1} \lambda x_1. h_{1,2}(x_0, x_1)(x_2)\ h_1(x_0)\ d\nu_1,$$

for all $x_0 \in X_0$, and a regular conditional density (with respect to ν_1) $h_{2,1} : X_0 \times X_2 \to \mathcal{D}(X_1)$ of f_1 given (f_0, f_2) defined by

$$h_{2,1}(x_0, x_2) = \frac{\lambda x_1. h_{1,2}(x_0, x_1)(x_2)\ h_1(x_0)}{\int_{X_1} \lambda x_1. h_{1,2}(x_0, x_1)(x_2)\ h_1(x_0)\ d\nu_1},$$

for all $x_0 \in X_0$ and $x_2 \in X_2$, such that, for all $x_1 \in X_1$ and $x_2 \in X_2$,

$$\lambda x_0.(h_2 \otimes h_{2,1})(x_0)(x_2, x_1) = \lambda x_0.(h_1 \otimes h_{1,2})(x_0)(x_1, x_2).$$

Proof. Note that $\lambda x_1. h_{1,2}(x_0, x_1)(x_2)\ h_1(x_0)$ can be written as $\lambda x_1. h_1(x_0)(x_1)\, h_{1,2}(x_0, x_1)(x_2)$. Define the conditional density $h : X_0 \to \mathcal{D}(X_2 \times X_1)$ by

$$h(x_0)(x_2, x_1) = (h_1 \otimes h_{1,2})(x_0)(x_1, x_2),$$

for all $x_0 \in X_0$, $x_1 \in X_1$, and $x_2 \in X_2$. By Proposition A.12.5, h_2 and $h_{2,1}$ are conditional densities and $h = h_2 \otimes h_{2,1}$. Hence, for all $x_1 \in X_1$ and $x_2 \in X_2$,

$$\lambda x_0.(h_2 \otimes h_{2,1})(x_0)(x_2, x_1) = \lambda x_0.(h_1 \otimes h_{1,2})(x_0)(x_1, x_2).$$

By Proposition A.12.6, h is a regular conditional density (with respect to $\nu_2 \otimes \nu_1$) of (f_2, f_1) given f_0. Then, again by Proposition A.12.6, h_2 is a regular conditional density (with respect to ν_2) of f_2 given f_0 and $h_{2,1}$ is a regular conditional density (with respect to ν_1) of f_1 given (f_0, f_2). □

Definition A.12.2. With Bayes theorem for conditional densities in the form

$$\lambda x_0.(h_2 \otimes h_{2,1})(x_0)(x_2, x_1) = \lambda x_0.(h_1 \otimes h_{1,2})(x_0)(x_1, x_2).$$

h_1 is the *prior*, $h_{1,2}$ is the *likelihood*, and $h_{2,1}$ is the *posterior*.

Now the issue of isolating the posterior $\mu_{2,1}$ in Bayes theorem for probability kernels (Proposition A.7.14) is discussed. (Recall the definition of projective product in Definition A.2.14.)

Proposition A.12.8. *(Bayes posterior) Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, $(X_1, \mathcal{A}_1, \nu_1)$ and $(X_2, \mathcal{A}_2, \nu_2)$ σ-finite measure spaces, and $f_0 : \Omega \to X_0$, $f_1 : \Omega \to X_1$, and $f_2 : \Omega \to X_2$ random variables. Suppose that $h_1 : X_0 \to \mathcal{D}(X_1)$ is a regular conditional density (with respect to ν_1) of f_1 given f_0 and $h_{1,2} : X_0 \times X_1 \to \mathcal{D}(X_2)$ is a regular conditional density (with respect to ν_2) of f_2 given (f_0, f_1). Let $\mu_1 \triangleq h_1 \cdot \nu_1$ and $\mu_{1,2} \triangleq h_{1,2} \cdot \nu_2$. Then the posterior $\mu_{2,1} : X_0 \times X_2 \to \mathcal{P}(X_1)$ is given by*

$$\mu_{2,1} = \lambda(x_0, x_2).\lambda x_1.h_{1,2}(x_0, x_1)(x_2) * \lambda(x_0, x_2).\mu_1(x_0).$$

Furthermore, $\lambda(x_0, x_2).\lambda x_1.h_{1,2}(x_0, x_1)(x_2) * \lambda(x_0, x_2).\mu_1(x_0)$ is a regular conditional distribution of f_1 given (f_0, f_2).

Proof. Note first that

$$\lambda(x_0, x_2).\lambda x_1.h_{1,2}(x_0, x_1)(x_2) : X_0 \times X_2 \to X_1 \to \mathbb{R}$$

has the property that the function from $X_0 \times X_2 \times X_1$ to $\mathbb{R}$ defined by $(x_0, x_2, x_1) \mapsto h_{1,2}(x_0, x_1)(x_2)$, for all $x_0 \in X_0$, $x_2 \in X_2$, and $x_1 \in X_1$, is measurable. Also

$$\lambda(x_0, x_2).\mu_1(x_0) : X_0 \times X_2 \to \mathcal{P}(X_1)$$

is a probability kernel. Thus the projective product in the definition of $\mu_{2,1}$ is well-defined.

Define $h_2 : X_0 \to \mathcal{D}(X_2)$ by

$$h_2(x_0) = \lambda x_2. \int_{X_1} \lambda x_1.h_{1,2}(x_0, x_1)(x_2)\ h_1(x_0)\ d\nu_1,$$

for all $x_0 \in X_0$, and $h_{2,1} : X_0 \times X_2 \to \mathcal{D}(X_1)$ by

$$h_{2,1}(x_0, x_2) = \frac{\lambda x_1.h_{1,2}(x_0, x_1)(x_2)\ h_1(x_0)}{\int_{X_1} \lambda x_1.h_{1,2}(x_0, x_1)(x_2)\ h_1(x_0)\ d\nu_1},$$

for all $x_0 \in X_0$ and $x_2 \in X_2$. Then, for all $x_0 \in X_0$ and $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$,

$$\begin{aligned}
&\lambda x_0.((h_2 \cdot \nu_2) \otimes (h_{2,1} \cdot \nu_1))(x_0)(A^*) \\
&= \lambda x_0.((h_2 \otimes h_{2,1}) \cdot (\nu_2 \otimes \nu_1))(x_0)(A^*) && \text{[Proposition A.12.3]} \\
&= \lambda x_0. \int_{X_2 \times X_1} \mathbf{1}_{A^*}\ (h_2 \otimes h_{2,1})(x_0)\ d(\nu_2 \otimes \nu_1) \\
&= \lambda x_0. \int_{X_1 \times X_2} \mathbf{1}_{A}\ (h_1 \otimes h_{1,2})(x_0)\ d(\nu_1 \otimes \nu_2) && \text{[Proposition A.12.7]} \\
&= \lambda x_0.((h_1 \otimes h_{1,2}) \cdot (\nu_1 \otimes \nu_2))(x_0)(A) \\
&= \lambda x_0.((h_1 \cdot \nu_1) \otimes (h_{1,2} \cdot \nu_2))(x_0)(A) && \text{[Proposition A.12.3]} \\
&= \lambda x_0.(\mu_1 \otimes \mu_{1,2})(x_0)(A).
\end{aligned}$$

Thus the posterior $\mu_{2,1}$ is given by $\mu_{2,1} = h_{2,1} \cdot \nu_1$.

Now, for all $x_0 \in X_0$, $x_2 \in X_2$, and $A_1 \in \mathcal{A}_1$,

$$\begin{aligned}
&(\lambda(x_0,x_2).\lambda x_1.h_{1,2}(x_0,x_1)(x_2) * \lambda(x_0,x_2).\mu_1(x_0))(x_0,x_2)(A_1) \\
&= \frac{\int_{X_1} \mathbf{1}_{A_1} \lambda x_1.h_{1,2}(x_0,x_1)(x_2)\; d\mu_1(x_0)}{\int_{X_1} \lambda x_1.h_{1,2}(x_0,x_1)(x_2)\; d\mu_1(x_0)} && \text{[Definition A.2.14]} \\
&= \frac{\int_{X_1} \mathbf{1}_{A_1} \lambda x_1.h_{1,2}(x_0,x_1)(x_2)\; h_1(x_0)\; d\nu_1}{\int_{X_1} \lambda x_1.h_{1,2}(x_0,x_1)(x_2)\; h_1(x_0)\; d\nu_1} && \text{[Proposition A.3.6]} \\
&= \int_{X_1} \mathbf{1}_{A_1} h_{2,1}(x_0,x_2)\; d\nu_1 \\
&= \int_{X_1} \mathbf{1}_{A_1}\; d(h_{2,1} \cdot \nu_1)(x_0,x_2) && \text{[Proposition A.3.6]} \\
&= (h_{2,1} \cdot \nu_1)(x_0,x_2)(A_1).
\end{aligned}$$

Hence $\lambda(x_0,x_2).\lambda x_1.h_{1,2}(x_0,x_1)(x_2) * \lambda(x_0,x_2).\mu_1(x_0) = h_{2,1} \cdot \nu_1$. By Proposition A.12.7, $h_{2,1}$ is a regular conditional density of f_1 given (f_0, f_2) and so $\lambda(x_0,x_2).\lambda x_1.h_{1,2}(x_0,x_1)(x_2) * \lambda(x_0,x_2).\mu_1(x_0)$ is a regular conditional distribution of f_1 given (f_0, f_2). □

Next comes a key factorization result. (Compare this result with Proposition A.7.19.)

Proposition A.12.9. *Let $(X_0, \mathcal{A}_0)$ be a measurable space, $(X_i, \mathcal{A}_i, \nu_i)$ a σ-finite measure space, for $i = 1, \ldots, n$, and $\mu : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$ a probability kernel. Suppose there exists a conditional density $h : X_0 \to \mathcal{D}(\prod_{i=1}^n X_i)$ such that $\mu = h \cdot \bigotimes_{i=1}^n \nu_i$. Then, for $i = 1, \ldots, n$, there exists a conditional density $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ such that $\mu = \bigotimes_{i=1}^n (h_i \cdot \nu_i)$.*

Proof. According to Proposition A.12.5, there exist conditional densities $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ such that $h = \bigotimes_{i=1}^n h_i$. Then $\mu = h \cdot \bigotimes_{i=1}^n \nu_i = (\bigotimes_{i=1}^n h_i) \cdot (\bigotimes_{i=1}^n \nu_i) = \bigotimes_{i=1}^n (h_i \cdot \nu_i)$, by Proposition A.12.3. □

Proposition A.12.10. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, $(X_i, \mathcal{A}_i, \nu_i)$ a σ-finite measure space, for $i = 1, \ldots, n$, and $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{i=1}^n X_i$ measurable functions. Suppose that $h : X_0 \to \mathcal{D}(\prod_{i=1}^n X_i)$ is a conditional density such that*

$$\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(h \cdot \bigotimes_{i=1}^n \nu_i)(f_0(\omega))(A) \text{ a.s.}$$

for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$. Then, for $i = 1, \ldots, n$, there exists a conditional density $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ such that

$$\mathsf{P}(f_i^{-1}(A_i) \mid (f_0, \ldots, f_{i-1})) = \lambda\omega.(h_i \cdot \nu_i)((f_0, \ldots, f_{i-1})(\omega))(A_i) \text{ a.s.}$$

for all $A_i \in \mathcal{A}_i$.

Proof. For $i = 1, \ldots, n$, define $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ according to Proposition A.12.9. The result then follows directly from Proposition A.7.16. □

Now the converse of Proposition A.12.10 is considered.

Proposition A.12.11. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, $(X_i, \mathcal{A}_i, \nu_i)$ a σ-finite measure space, for $i = 1, \ldots, n$, and $f_0 : \Omega \to X_0$ and $f : \Omega \to \prod_{i=1}^n X_i$ measurable functions. Suppose that, for $i = 1, \ldots, n$, there exists a conditional density $h_i : \prod_{j=0}^{i-1} X_j \to \mathcal{D}(X_i)$ such that*

$$\mathsf{P}(f_i^{-1}(A_i) \mid (f_0, \ldots, f_{i-1})) = \lambda\omega.(h_i \cdot \nu_i)((f_0, \ldots, f_{i-1})(\omega))(A_i) \text{ a.s.}$$

for all $A_i \in \mathcal{A}_i$. Then there exists a conditional density $h : X_0 \to \mathcal{D}(\prod_{i=1}^n X_i)$ such that

$$\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(h \cdot \bigotimes_{i=1}^n \nu_i)(f_0(\omega))(A) \text{ a.s.}$$

for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$.

Proof. Let $h \triangleq \bigotimes_{i=1}^n h_i$. The result then follows directly from Proposition A.7.12, using Proposition A.12.3. □

In other words, together Propositions A.12.10 and A.12.11 state that $h : X_0 \to \mathcal{D}(\prod_{i=1}^n X_i)$ is a regular conditional density (with respect to $\bigotimes_{i=1}^n \nu_i$) if and only if, for $i = 1, \ldots, n$, h_i is a regular conditional density (with respect to ν_i).

Here is the fundamental result concerning the deconstruction of a probability kernel mapping into probability measures on a product space, under the assumption that the probability kernel is defined by a conditional density.

Proposition A.12.12. *Let $(\Omega, \mathfrak{S}, \mathsf{P})$ be a probability space, $(X_0, \mathcal{A}_0)$ a measurable space, $(X_i, \mathcal{A}_i)$ a standard Borel space, for $i = 1, \ldots, n$, and $f_0 : \Omega \to X_0$ and $f_i : \Omega \to X_i$ measurable functions, for $i = 1, \ldots, n$. Suppose there is a dependency graph with vertices $0, \ldots, n$, where vertex i is labelled by $\sigma(f_i)$, for $i = 0, \ldots, n$, such that $0, \ldots, n$ is a topological order of the vertices and*

$$\sigma(f_i) \underset{\sigma(f_0, f_{par(i)})}{\perp\!\!\!\perp} \sigma(f_0, \ldots, f_{i-1}),$$

for $i = 1, \ldots, n$. Let $\mu : X_0 \to \mathcal{P}(\prod_{i=1}^n X_i)$ be a regular conditional distribution of $(f_1, \ldots, f_n)$ given f_0. Let ν_i be a σ-finite measure on X_i, for $i = 1, \ldots, n$, and suppose that there exists a conditional density $h : X_0 \to \mathcal{D}(\prod_{i=1}^n X_i)$ such that $\mu = h \cdot \bigotimes_{i=1}^n \nu_i$. Then there exist conditional densities $h_i : X_0 \times \prod_{j \in par(i)} X_j \to \mathcal{D}(X_i)$, for $i = 1, \ldots, n$, such that

$$\mu = \bigotimes_{i=1}^n \lambda(x_0, \ldots, x_{i-1}).h_i(x_0, x_{par(i)}) \cdot \nu_i) \ \ \mathcal{L}(f_0)\text{-a.e.}$$

Proof. Since $\sigma(f_i) \perp\!\!\!\perp_{\sigma(f_0, f_{par(i)})} \sigma(f_0, \ldots, f_{i-1})$, it follows from Proposition A.6.1 that, for $i = 1, \ldots, n$,

$$\mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, \ldots, f_{i-1})) = \mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, f_{par(i)})) \text{ a.s.},$$

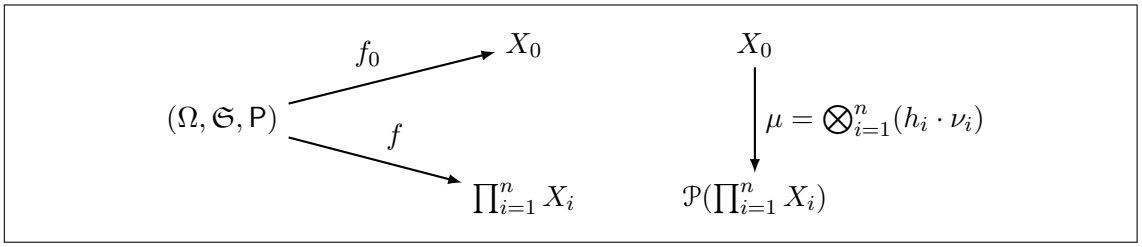

Figure A.42: Setting for Proposition A.12.12

for all $A_i \in \mathcal{A}_i$.

By Proposition TO BE ADDED, for $i = 1, \ldots, n$, there exists a conditional density $h_i : X_0 \times \prod_{j \in par(i)} X_j \to \mathcal{D}(X_i)$ such that

$$\mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, f_{par(i)})) = \lambda\omega.(h_i \cdot \nu_i)((f_0, f_{par(i)})(\omega))(A_i) \text{ a.s.},$$

for all $A_i \in \mathcal{A}_i$. Hence, for $i = 1, \ldots, n$, and for all $A_i \in \mathcal{A}_i$, almost surely,

$$\begin{aligned}
&\mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, \ldots, f_{i-1})) \\
&= \mathsf{P}(f_i^{-1}(A_i) \,|\, (f_0, f_{par(i)})) \\
&= \lambda\omega.(h_i \cdot \nu_i)((f_0, f_{par(i)})(\omega))(A_i) \\
&= \lambda\omega.\lambda(x_0, \ldots, x_{i-1}).(h_i \cdot \nu_i)(x_0, x_{par(i)})((f_0, \ldots, f_{i-1})(\omega))(A_i).
\end{aligned}$$

By Proposition A.7.12,

$$\mathsf{P}(f^{-1}(A) \mid f_0) = \lambda\omega.(\bigotimes_{i=1}^n \lambda(x_0, \ldots, x_{i-1}).(h_i \cdot \nu_i)(x_0, x_{par(i)}))(f_0(\omega))(A) \text{ a.s.},$$

for all $A \in \bigotimes_{i=1}^n \mathcal{A}_i$. Thus, by the uniqueness part of Proposition A.5.16, it follows that

$$\mu = \bigotimes_{i=1}^n \lambda(x_0, \ldots, x_{i-1}).(h_i \cdot \nu_i)(x_0, x_{par(i)}) \;\; \mathcal{L}(f_0)\text{-a.e.},$$

that is,

$$\mu = \bigotimes_{i=1}^n (\lambda(x_0, \ldots, x_{i-1}).h_i(x_0, x_{par(i)}) \cdot \nu_i) \;\; \mathcal{L}(f_0)\text{-a.e.}.$$

□

A.13 Sums of Conditional Densities

Next, sums of conditional densities are considered.

Let I be a countable index set, and $(X_i, \mathcal{A}_i, \nu_i)$ a measure space and h_i a density on $(X_i, \mathcal{A}_i, \nu_i)$, for all $i \in I$. Suppose $g : I \to [0, \infty)$ is a function such that $\sum_{i \in I} g(i) = 1$. Now define

$$g \bullet \bigoplus_{i \in I} h_i : \coprod_{i \in I} X_i \to \mathbb{R}$$

by

$$(g \bullet \bigoplus_{i\in I} h_i)(x) = g(i)h_i(x),$$

whenever $x \in X_i$. Then $g \bullet \bigoplus_{i\in I} h_i$ is measurable, $g \bullet \bigoplus_{i\in I} h_i \geq 0$, and

$$\begin{aligned}
&\int_{\coprod_{i\in I} X_i} (g \bullet \bigoplus_{i\in I} h_i)\, d\bigoplus_{i\in I} \nu_i \\
&= \sum_{i\in I} \int_{X_i} g(i)h_i \, d\mu_i && [\textstyle\bigoplus_{i\in I} \nu_i \text{ restricted to } X_i \text{ is } \nu_i] \\
&= \sum_{i\in I} g(i) && [\text{Each } h_i \text{ is a density}] \\
&= 1.
\end{aligned}$$

Thus $g \bullet \bigoplus_{i\in I} h_i$ is a density on $(\coprod_{i\in I} X_i, \bigoplus_{i\in I} \mathcal{A}_i, \bigoplus_{i\in I} \nu_i)$.

More generally, let $(X, \mathcal{A})$ be a measurable space, $h_i : X \to \mathcal{D}(X_i)$ a conditional density, for $i \in I$, and $g : X \to I \to [0, \infty)$ a weight function. Define

$$g \bullet \bigoplus_{i\in I} h_i : X \to \coprod_{i\in I} X_i \to \mathbb{R}$$

by

$$(g \bullet \bigoplus_{i\in I} h_i)(x)|_{X_i} = g(x)(i)h_i(x),$$

for all $i \in I$ and $x \in X$. Then $g \bullet \bigoplus_{i\in I} h_i$ is a conditional density.

Proposition A.13.1. *Let $(X, \mathcal{A})$ be a measurable space, $(X_i, \mathcal{A}_i, \nu_i)$ a measure space, $h_i : X \to \mathcal{D}(X_i)$ a conditional density, for all $i \in I$, and $g : X \to I \to [0, \infty)$ a weight function. Then*

$$g \bullet \bigoplus_{i\in I} h_i : X \to \mathcal{D}(\coprod_{i\in I} X_i)$$

is a conditional density.

Proof. By the remarks above, $(g \bullet \bigoplus_{i\in I} h_i)(x) \in \mathcal{D}(\coprod_{i\in I} X_i)$, for all $x \in X$. Furthermore, for all $i \in I$, the function defined by $(x, x_i) \mapsto (g \bullet \bigoplus_{i\in I} h_i)(x)(x_i)$, for all $x \in X$ and $x_i \in X_i$, is measurable. Thus $g \bullet \bigoplus_{i\in I} h_i$ is a conditional density. □

Now a converse of Proposition A.13.1 is considered.

Proposition A.13.2. *Let $(X, \mathcal{A})$ be a measurable space, $(X_i, \mathcal{A}_i, \nu_i)$ a σ-finite measure space, for $i \in I$, and $\mu : X \to \mathcal{P}(\coprod_{i\in I} X_i)$ a probability kernel. Suppose there exists a conditional density $h : X \to \mathcal{D}(\coprod_{i\in I} X_i)$ such that $\mu = h \cdot \bigoplus_{i\in I} \nu_i$ and, for all $i \in I$ and $x \in X$, $\int_{X_i} h(x)|_{X_i} \, d\nu_i \neq 0$. Then there exist a weight function $g : X \to I \to [0, \infty)$, and, for $i \in I$, a conditional density $h_i : X \to \mathcal{D}(X_i)$ such that*

$$h = g \bullet \bigoplus_{i\in I} h_i \text{ and } \mu = g \bullet \bigoplus_{i\in I} (h_i \cdot \nu_i).$$

Proof. Define $g : X \to I \to [0, \infty)$ by

$$g(x)(i) = \int_{\coprod_{i \in I} X_i} \mathbf{1}_{X_i} h(x) \, d \bigoplus_{i \in I} \nu_i,$$

for all $x \in X$ and $i \in I$. By Proposition A.2.15, $\lambda x.g(x)(i)$ is measurable, for all $i \in I$. Also, for all $x \in X$,

$$\begin{aligned}
&\sum_{i \in I} g(x)(i) \\
&= \sum_{i \in I} \int_{\coprod_{i \in I} X_i} \mathbf{1}_{X_i} h(x) \, d \bigoplus_{i \in I} \nu_i \\
&= \int_{\coprod_{i \in I} X_i} (\sum_{i \in I} \mathbf{1}_{X_i}) h(x) \, d \bigoplus_{i \in I} \nu_i \\
&= \int_{\coprod_{i \in I} X_i} h(x) \, d \bigoplus_{i \in I} \nu_i \\
&= 1.
\end{aligned}$$

Thus g is a weight function.

For all $i \in I$, define $h_i : X \to \mathcal{D}(X_i)$ by

$$h_i(x) = \frac{h(x)|_{X_i}}{\int_{X_i} h(x)|_{X_i} \, d\nu_i},$$

for all $x \in X$. Then, for all $i \in I$,

$$\begin{aligned}
&\int_{X_i} h_i(x) \, d\nu_i \\
&= \int_{X_i} \frac{h(x)|_{X_i}}{\int_{X_i} h(x)|_{X_i} \, d\nu_i} \, d\nu_i \\
&= 1.
\end{aligned}$$

Furthermore, for all $i \in I$, the function defined by $(x, x) \mapsto h_i(x)(x)$, for all $x \in X$ and $x \in X_i$, is measurable. Hence, for all $i \in I$, h_i is a conditional density.

Whenever $x_i \in X_i$,

$$\begin{aligned}
&(g \bullet \bigoplus_{i \in I} h_i)(x)(x_i) \\
&= g(x)(i) h_i(x)(x_i) \\
&= \left(\int_{\coprod_{i \in I} X_i} \mathbf{1}_{X_i} h(x) \, d \bigoplus_{i \in I} \nu_i \right) \left(\frac{h(x)|_{X_i}(x_i)}{\int_{X_i} h(x)|_{X_i} \, d\nu_i} \right) \\
&= h(x)|_{X_i}(x_i).
\end{aligned}$$

Hence $h = g \bullet \bigoplus_{i \in I} h_i$.

For all $x \in X$ and, for all $i \in I$, $A_i \in \mathcal{A}_i$,

$$\begin{aligned}
&\mu(x)(\coprod_{i\in I} A_i) \\
&= (h \cdot \bigoplus_{i\in I} \nu_i)(x)(\coprod_{i\in I} A_i) \\
&= ((g \bullet \bigoplus_{i\in I} h_i) \cdot \bigoplus_{i\in I} \nu_i)(x)(\coprod_{i\in I} A_i) \\
&= \int_{\coprod_{i\in I} X_i} \mathbf{1}_{\coprod_{i\in I} A_i} (g \bullet \bigoplus_{i\in I} h_i)(x)\ d \bigoplus_{i\in I} \nu_i \\
&= \sum_{i\in I} \int_{X_i} \mathbf{1}_{A_i} (g \bullet \bigoplus_{i\in I} h_i)(x)|_{X_i}\ d\nu_i \\
&= \sum_{i\in I} \int_{X_i} \mathbf{1}_{A_i}\, g(x)(i) h_i(x)\ d\nu_i \\
&= \sum_{i\in I} g(x)(i) \left(\int_{X_i} \mathbf{1}_{A_i} h_i(x)\ d\nu_i \right) \\
&= \sum_{i\in I} g(x)(i)(h_i \cdot \nu_i)(x)(A_i) \\
&= (g \bullet \bigoplus_{i\in I} (h_i \cdot \nu_i))(x)(\coprod_{i\in I} A_i)
\end{aligned}$$

Hence $\mu = g \bullet \bigoplus_{i\in I}(h_i \cdot \nu_i)$. □

A.14 Quotients of Conditional Densities

To do: Add here the analogue of Section A.10 for conditional densities.

A.15 Restrictions of Conditional Densities

To do: Add here the analogue of Section A.11 for conditional densities.

A.16 Computing Integrals

There are a variety of computing tasks that an agent must perform during deployment. Here is an (incomplete) list of such tasks.

1. *Compute expected utility*
   Let $f : Y \to \mathbb{R}$ be a utility function and $\lambda x.\mu_n(h_n, x) : X \to \mathcal{P}(Y)$ an empirical belief. Then the expected utility at some $x \in X$ is given by

$$\int_Y f\ d\mu_n(h_n, x).$$

2. *Compute the fusion of probability kernels*
   Let $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels. The fusion $\mu_1 \odot \mu_2 : X_0 \to \mathcal{P}(X_2)$ of μ_1 and μ_2 is defined by

$$(\mu_1 \odot \mu_2)(x_0) = \lambda A_2. \int_{X_1} \lambda x_1.\mu_2(x_0, x_1)(A_2)\ d\mu_1(x_0),$$

   for all $x_0 \in X_0$.

3. *Compute the product of probability kernels*
   Let $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels. The product $\mu_1 \otimes \mu_2 : X_0 \to \mathcal{P}(X_1 \times X_2)$ of μ_1 and μ_2 is defined by

$$(\mu_1 \otimes \mu_2)(x_0) = \lambda A. \int_{X_1} \left(\lambda x_1. \int_{X_2} \lambda x_2.\mathbf{1}_A(x_1, x_2)\ d\mu_2(x_0, x_1) \right)\ d\mu_1(x_0),$$

   for all $x_0 \in X_0$.

4. *Compute the projective product of a function and a probability kernel*
   Let $f : X \to Y \to \mathbb{R}$ a non-negative function such that $\lambda(x, y).f(x)(y) : X \times Y \to \mathbb{R}$ is measurable, and $\mu : X \to \mathcal{P}(Y)$ a probability kernel such that $0 < \int_Y f(x)\ d\mu(x) < \infty$, for all $x \in X$. The projective product $f * \mu : X \to \mathcal{P}(Y)$ of f and μ is defined by

$$(f * \mu)(x)(B) = \frac{\int_Y \mathbf{1}_B f(x)\ d\mu(x)}{\int_Y f(x)\ d\mu(x)},$$

   for all $x \in X$ and $B \in \mathcal{B}$.

5. *Filter schemas and empirical beliefs*
   The filter recurrence equation for empirical beliefs (in conditional density form) is as follows.

$$\begin{aligned}
&\lambda x.f_{\mu_{n+1}}(h_{n+1}, x) = \\
&\lambda x.\lambda y.f_{\xi_{n+1}}(h_n, a_{n+1}, x, y)(o_{n+1}) * \\
&\quad \lambda x.((f_{\nu_n}(h_n) \otimes \lambda x'.f_{\mu_n}(h_n, x')) \odot \lambda(x', y').f_{\tau_{n+1}}(h_n, a_{n+1}, x, x', y')).
\end{aligned}$$

The common thread for all these tasks is that they all involve computing integrals. Now integrals are defined as limits. Thus, except in the few cases where symbolic integration is possible, integrals can only be computed approximately, by Monte Carlo methods, for example.

Task 1 is comparatively straightforward: an expected utility is a real number which can be accurately approximated by standard Monte Carlo methods. For this, let $\mu : \mathcal{P}(X)$ and $f : X \to \mathbb{R}$ be integrable, then

$$\int_X f\ d\mu \approx \frac{1}{M} \sum_{i=1}^{M} f(x^{(i)}),$$

where $x^{(1)}, \ldots, x^{(M)}$ are iid samples from μ. The approximation is justified by the law of large numbers.

If $\mu_1 : \mathcal{P}(X_1)$ and $\mu_2 : X_1 \to \mathcal{P}(X_2)$ is a probability kernel, then to sample from $\mu_1 \odot \mu_2 : \mathcal{P}(X_2)$, one first samples from μ_1 to get, say, $x_1 \in X_1$, and then samples from $\mu_2(x_1)$ to get, say, $x_2 \in X_2$. Then x_2 is the sample from $\mu_1 \odot \mu_2$. (The sample x_1 is discarded.) Similarly, to sample from $\mu_1 \otimes \mu_2 : \mathcal{P}(X_1 \times X_2)$, one first samples from μ_1 to get, say, $x_1 \in X_1$, and then samples from $\mu_2(x_1)$ to get, say, $x_2 \in X_2$. Then $(x_1, x_2) \in X_1 \times X_2$ is the sample from $\mu_1 \otimes \mu_2$. Each of these is ancestral sampling, where, for fusion, the first coordinate is discarded. Both sampling methods can be generalized in the obvious way to a fusion with n components and a product with n components, where $n \geq 2$.

Thus, if $f : X_2 \to \mathbb{R}$ is integrable, then

$$\int_X f \, d(\mu_1 \odot \mu_2) \approx \frac{1}{M} \sum_{i=1}^{M} f(x^{(i)}),$$

where $x^{(1)}, \ldots, x^{(M)}$ are iid samples from $\mu_1 \odot \mu_2$. Similarly, if $f : X_1 \otimes X_2 \to \mathbb{R}$ is integrable, then

$$\int_{X_1 \times X_2} f \, d(\mu_1 \otimes \mu_2) \approx \frac{1}{M} \sum_{i=1}^{M} f(x_1^{(i)}, x_2^{(i)}),$$

where $(x_1^{(1)}, x_2^{(1)}) \ldots, (x_1^{(M)}, x_2^{(M)})$ are iid samples from $\mu_1 \otimes \mu_2$.

Let $f : Y \to \mathbb{R}$ be a non-negative measurable function and $\mu : \mathcal{P}(Y)$ such that $0 < \int_Y f \, d\mu < \infty$. Let $g : Y \to \mathbb{R}$ be a function that is integrable with respect to $f * \mu$. Then, by Proposition A.2.20,

$$\int_Y g \, d(f * \mu) = \frac{\int_Y gf \, d\mu}{\int_Y f \, d\mu}.$$

Thus

$$\int_Y g \, d(f * \mu) \approx \frac{\sum_{i=1}^{M} g(y^{(i)}) f(y^{(i)})}{\sum_{i=1}^{M} f(y^{(i)})},$$

where $y^{(1)}, \ldots, y^{(M)}$ are iid samples from μ.

In contrast, Tasks 2 to 5 are much more problematic, and Task 5 even includes as subtasks each of Tasks 2 to 4. In Section 4.2, an approach to handling Task 5 is outlined. Clearly, similar ideas can be used to handle Tasks 2 to 4. In any case, an agent needs specialized software support for handling each of these tasks.

To do: Complete this section.

Bibliographical Notes

This chapter has presented mostly standard concepts and results of probability theory concentrating on the notions of probability kernel and regular conditional distribution that are central to this book. An excellent introductory book on probability that is concise and rigorous is [81]. More detail can be found, for example, in [13, 24, 43, 83, 87]. In addition, [144, Chapter 2] provides an intuitive, well-motivated introduction to the central concept of conditional expectation.

Exercises

A.1 Let $(X_0, \mathcal{A}_0)$, $(X_1, \mathcal{A}_1)$ and $(X_2, \mathcal{A}_2)$ be measurable spaces, and $\mu_1 : X_0 \to \mathcal{P}(X_1)$ and $\mu_2 : X_0 \times X_1 \to \mathcal{P}(X_2)$ probability kernels. Suppose that X_1 is a singleton set. Prove that $\mu_1 \otimes \mu_2 : X_0 \to \mathcal{P}(X_1 \times X_2)$ can be identified with μ_2.

A.2 Prove Proposition A.7.3.

A.3 Prove Proposition A.7.5.

A.4 Prove Proposition A.7.7.

A.5 Let $(X_i, \mathcal{A}_i)$ be a measurable space and $\mu_i : \prod_{j=1}^{i-1} X_j \to \mathcal{P}(X_i)$ a probability kernel, for $i = 1, \ldots, n$. Let $f : \prod_{i=1}^{n} X_i \to \mathbb{R}$ be a non-negative measurable function such that, for all $i = 1, \ldots, n$ and $x_1 \in X_1, \ldots, x_{i-1} \in X_{i-1}$,

$$0 < \int_{\prod_{j=i}^{n} X_j} \lambda(x_i, \ldots, x_n).f(x_1, \ldots, x_n)\ d(\bigotimes_{j=i}^{n} \mu_j)(x_1, \ldots, x_{i-1}) < \infty.$$

For $i = 1, \ldots, n$, define

$$f_i : \prod_{j=1}^{i-1} X_j \to X_i \to \mathbb{R}$$

by

$$f_i(x_1, \ldots, x_{i-1})(x_i) = \int_{\prod_{j=i+1}^{n} X_j} \lambda(x_{i+1}, \ldots, x_n).f(x_1, \ldots, x_n)\ d(\bigotimes_{j=i+1}^{n} \mu_j)(x_1, \ldots, x_i),$$

for all $x_1 \in X_1, \ldots, x_i \in X_i$. Prove that, for all $i = 1, \ldots, n$, the function

$$\lambda(x_1, \ldots, x_i).f_i(x_1, \ldots, x_{i-1})(x_i) : \prod_{j=1}^{i} X_j \to \mathbb{R}$$

is measurable. Furthermore, prove that

$$f * (\bigotimes_{i=1}^{n} \mu_i) = \bigotimes_{i=1}^{n} (f_i * \mu_i).$$

A.6 For all $n \in \mathbb{N}$, let $\pi_{1,\ldots,n} : \prod_{n\in\mathbb{N}} X_n \to \prod_{i=1}^{n} X_i$ be the canonical projection. Let $(X_n, \mathcal{A}_n)$ be a measurable space and $\mu_n : \prod_{j=1}^{n-1} X_j \to \mathcal{P}(X_n)$ a probability kernel, for all $n \in \mathbb{N}$. Let $f : \prod_{n\in\mathbb{N}} X_n \to \mathbb{R}$ be a non-negative measurable function such that $0 < \int_{\prod_{n\in\mathbb{N}} X_n} f\ d \bigotimes_{n\in\mathbb{N}} \mu_n < \infty$. For all $n \in \mathbb{N}$, define $g_n : \prod_{i=1}^{n} X_i \to \mathbb{R}$ by

$$g_n(x_1, \ldots, x_n) = \int_{\prod_{j=n+1}^{\infty} X_j} \lambda(x_{n+1}, x_{n+2}, \ldots).f(x_1, x_2, \ldots)\ d(\bigotimes_{j=n+1}^{\infty} \mu_j)(x_1, \ldots, x_n),$$

for all $x_1 \in X_1, \ldots, x_n \in X_n$. Prove that, for all $n \in \mathbb{N}$, g_n is a non-negative measurable function such that $0 < \int_{\prod_{i=1}^n X_i} g_n \, d \bigotimes_{i=1}^n \mu_i < \infty$. Furthermore, prove that, for all $n \in \mathbb{N}$,

$$(f * \bigotimes_{n \in \mathbb{N}} \mu_n) \circ \pi_{1,\ldots,n}^{-1} = g_n * (\bigotimes_{i=1}^n \mu_i).$$

A.7 Let $(X_n, \mathcal{A}_n)$ be a measurable space and $\mu_n : \prod_{j=1}^{n-1} X_j \to \mathcal{P}(X_n)$ a probability kernel, for all $n \in \mathbb{N}$. Let $f : \prod_{n \in \mathbb{N}} X_n \to \mathbb{R}$ be a non-negative measurable function such that, for all $n \in \mathbb{N}$ and $x_1 \in X_1, \ldots, x_{n-1} \in X_{n-1}$,

$$0 < \int_{\prod_{j=n}^\infty X_j} \lambda(x_n, x_{n+1}, \ldots).f(x_1, x_2, \ldots) \, d(\bigotimes_{j=n}^\infty \mu_j)(x_1, \ldots, x_{n-1}) < \infty.$$

For all $n \in \mathbb{N}$, define $g_n : \prod_{i=1}^n X_i \to \mathbb{R}$ by

$$g_n(x_1, \ldots, x_n) = \int_{\prod_{j=n+1}^\infty X_j} \lambda(x_{n+1}, x_{n+2}, \ldots).f(x_1, x_2, \ldots) \, d(\bigotimes_{j=n+1}^\infty \mu_j)(x_1, \ldots, x_n),$$

for all $x_1 \in X_1, \ldots, x_n \in X_n$. For all $n \in \mathbb{N}$, define $f_n : \prod_{j=1}^{n-1} X_j \to X_n \to \mathbb{R}$ by

$$f_n(x_1, \ldots, x_{n-1})(x_n) = g_n(x_1, \ldots, x_n),$$

for all $x_1 \in X_1, \ldots, x_n \in X_n$. Prove that, for all $n \in \mathbb{N}$, $\lambda(x_1, \ldots, x_n).f_n(x_1, \ldots, x_{n-1})(x_n) : \prod_{j=1}^n X_j \to \mathbb{R}$ is a non-negative, measurable function such that, for all $x_1 \in X_1, \ldots, x_{n-1} \in X_{n-1}$,

$$0 < \int_{X_n} f_n(x_1, \ldots, x_{n-1}) \, d\mu_n(x_1, \ldots, x_{n-1}) < \infty.$$

Furthermore, prove that, for all $n \in \mathbb{N}$,

$$g_n * (\bigotimes_{i=1}^n \mu_i) = \bigotimes_{i=1}^n (f_i * \mu_i).$$

A.8 Let $(X_n, \mathcal{A}_n)$ be a measurable space and $\mu_n : \prod_{j=1}^{n-1} X_j \to \mathcal{P}(X_n)$ a probability kernel, for all $n \in \mathbb{N}$, and $f : \prod_{n \in \mathbb{N}} X_n \to \mathbb{R}$ a non-negative measurable function such that $0 < \int_{\prod_{n \in \mathbb{N}} X_n} f \, d \bigotimes_{n \in \mathbb{N}} \mu_n < \infty$. For all $n \in \mathbb{N}$, define $f_n : \prod_{j=1}^{n-1} X_j \to X_n \to \mathbb{R}$ by

$$f_n(x_1, \ldots, x_{n-1})(x_n) = \int_{\prod_{j=n+1}^\infty X_j} \lambda(x_{n+1}, x_{n+2}, \ldots).f(x_1, x_2, \ldots) \, d(\bigotimes_{j=n+1}^\infty \mu_j)(x_1, \ldots, x_n),$$

for all $x_1 \in X_1, \ldots, x_n \in X_n$. Prove that, for all $n \in \mathbb{N}$,

$$(f * \bigotimes_{n \in \mathbb{N}} \mu_n) \circ \pi_{1,\ldots,n}^{-1} = \bigotimes_{i=1}^n (f_i * \mu_i).$$

Appendix B

Logic

THIS appendix gives an account of the syntax, semantics, and a reasoning system for modal higher-order logic. Well-founded sets are also discussed.

B.1 Syntax

This section contains an account of the syntactic objects of the logic, such as types, terms, substitutions, standard predicates, and predicate rewrite systems.

B.1.1 Types

Definition B.1.1. An *alphabet* consists of three (mutually disjoint) sets:

1. A set $\mathfrak{T}$ of type constructors.
2. A set $\mathfrak{C}$ of constants.
3. A set $\mathfrak{V}$ of variables.

Each type constructor in $\mathfrak{T}$ has an arity. A *nullary* type constructor is a type constructor of arity 0. The set $\mathfrak{T}$ always includes the nullary type constructor o, which is the type of the booleans. Each constant in $\mathfrak{C}$ has a signature (see below). Constants can be declared to be *rigid*. In the semantics introduced below, rigid constants have the same meaning in every world. The set $\mathfrak{V}$ is denumerable. Variables are typically denoted by $x, y, z, \ldots$. For any particular application, the alphabet is assumed fixed and all definitions are relative to the alphabet.

Types are built up from the set of type constructors, using the symbols $\to$ and $\times$.

Definition B.1.2. A *type* is defined inductively as follows.

1. If T is a type constructor of arity k and $\alpha_1, \ldots, \alpha_k$ are types, then $T\ \alpha_1 \ldots \alpha_k$ is a type. (For $k = 0$, this reduces to a type constructor of arity 0 being a type.)
2. If α and β are types, then $\alpha \to \beta$ is a type.
3. If $\alpha_1, \ldots, \alpha_n$ are types, then $\alpha_1 \times \cdots \times \alpha_n$ is a type.

Notation. $\mathfrak{S}$ denotes the set of all types obtained from an alphabet ($\mathfrak{S}$ for 'sort').

The symbol $\to$ is right associative, so that $\alpha \to \beta \to \gamma$ means $\alpha \to (\beta \to \gamma)$. Each variable has a type. It is assumed that $\mathfrak{V}$ contains infinitely many variables of each type and the sets of variables of each type are pairwise disjoint.

Example B.1.1. In practical applications of the logic, a variety of types is needed. For example, declarative programming languages typically admit the following types (which are nullary type constructors): *Bool* ($\triangleq o$), *Nat* (the type of natural numbers), *Int* (the type of integers), *Float* (the type of floating-point numbers), *Real* (the type of real numbers), *Char* (the type of characters), and *String* (the type of strings).

Other useful type constructors are those used to define lists, trees, and so on. In the logic, *List* denotes the (unary) list type constructor. Thus, if α is a type, then *List* α is the type of lists whose elements have type α.

A useful property of a type is its order.

Definition B.1.3. The *order* of a type is defined by induction as follows.

1. A type has order 0 if it is a type constructor of arity 0.
2. A type has order $\max\{order(\alpha_1), \ldots, order(\alpha_k)\}$, if it has the form $T\ \alpha_1 \ldots \alpha_k$, where $order(\alpha_i)$ is the order of α_i, for $i = 1, \ldots, k$.
3. A type has order $\max\{order(\alpha_1), \ldots, order(\alpha_n)\}$, if it has the form $\alpha_1 \times \cdots \times \alpha_n$, where $order(\alpha_i)$ is the order of α_i, for $i = 1, \ldots, n$.
4. A type has order $1 + \max\{order(\alpha), order(\beta)\}$, if it has the form $\alpha \to \beta$, where $order(\alpha)$ is the order of α and $order(\beta)$ is the order of β.

The order is defined for each type. For example, the order of $(\alpha \to \beta \to o) \times \gamma$ is 2.

For the purposes of this book, the rank of a type will be much more useful than its order.

Definition B.1.4. The *rank* of a type is defined by induction as follows.

1. A type has rank 0 if it has the form $T\ \alpha_1 \ldots \alpha_k$ or $\alpha_1 \times \cdots \times \alpha_n$.
2. A type has rank $n + 1$ if it has the form $\alpha \to \beta$, where β is a type of rank n.

The rank is defined for each type. A type has rank n if it has the form

$$\alpha_1 \to \cdots \to \alpha_n \to \beta,$$

where β has rank 0. Intuitively, the rank of a type is the number of 'top-level' arrows in the type. Note that the rank of a type is generally quite different to its order. For example, the rank of $(\alpha \to \beta \to o) \times \gamma$ is 0.

Much importance is traditionally attached to the class of terms that have type o, that is, the class of formulas. However, it will be shown that formulas can generally be replaced throughout by a bigger class of terms called biterms and that this generalisation is useful. For example, biterms can be arguments of the connectives and quantifiers. Moreover, the dual modalities will be introduced and shown to be applicable to biterms. Also theories can consist of biterms, not just formulas.

Next the type of biterms is introduced.

Definition B.1.5. A *biterm type of rank* n ($n \geq 0$) is a type having the form

$$\alpha_1 \to \cdots \to \alpha_n \to o.$$

A *biterm type* is a biterm type of rank n, for some $n \geq 0$.

There is just one biterm type of rank 0 and that is o. Note that, if $n \geq 1$, then $\alpha_2 \to \cdots \to \alpha_n \to o$ is a biterm type of rank $n-1$. Also, if α is any type, then $\alpha \to \alpha_1 \to \cdots \to \alpha_n \to o$ is a biterm type of rank $n+1$.

Notation. If $\alpha \triangleq \alpha_1 \to \cdots \to \alpha_n \to o$ is a biterm type of rank > 0, then $\mathsf{t}(\alpha)$, the *tail* of α, is defined by $\mathsf{t}(\alpha) = \alpha_2 \to \cdots \to \alpha_n \to o$.

Thus $\mathsf{t}(\alpha_1 \to o) = o$ and $\mathsf{t}(\alpha_1 \to \alpha_2 \to o) = \alpha_2 \to o$.

Definition B.1.6. A *signature* is the declared type for a constant.

The fact that a constant C has signature α is often denoted by $C : \alpha$.

Amongst the constants, the *data constructors* are distinguished. In a knowledge representation context, data constructors are used to represent individuals. In a programming language context, data constructors are used to construct data values. In contrast, constants that are not data constructors usually are used to compute with data values; such constants have definitions while data constructors do not. In the semantics for the logic, data constructors are used to construct models. As examples, each constant standing for an integer is a data constructor, and similarly for each floating-point number and character. The constants $\sharp_\alpha$ (cons) used to construct lists (whose items have type α) are data constructors. All data constructors are rigid constants. Data constructors always have a signature of the form $\sigma_1 \to \cdots \to \sigma_n \to (T\ \alpha_1 \ldots \alpha_k)$, where T is a type constructor of arity k. The *arity* of the data constructor is n. A *nullary* data constructor is a data constructor of arity 0.

In typical applications, many constants that are not data constructors are not rigid. For example, a constant whose definition is believed by agents may have a different definition for different agents and/or a different definition for the same agent at different times (and therefore its meaning may vary from world to world).

The set $\mathfrak{C}$ always includes the following rigid constants.

1. $=_\alpha$, having signature $\alpha \to \alpha \to o$, for each type α.
2. $\top_o$ and $\bot_o$, having signature o.
3. $\neg_\alpha$, having signature $\alpha \to \alpha$, for each biterm type α.
4. $\wedge_\alpha$, $\vee_\alpha$, $\longrightarrow_\alpha$, and $\longleftarrow_\alpha$, having signature $\alpha \to \alpha \to \alpha$, for each biterm type α.
5. Σ_α and Π_α, having signature $\alpha \to o$, for each biterm type α.

The semantics of these constants is given in Appendix B.2, but it is helpful to give an intuitive description of the intended meanings of these constants now. For each type α, there corresponds a domain $\mathcal{D}_\alpha$ containing the meanings of the terms of type α. The domain $\mathcal{D}_o$ for o is $\mathbb{B}$. For a type of the form $\alpha \to \beta$, the domain $\mathcal{D}_{\alpha\to\beta}$ is a collection of mappings from $\mathcal{D}_\alpha$ to $\mathcal{D}_\beta$.

The intended meaning of $=_\alpha$ is the function $\lambda x.\lambda y.(x = y) : \mathcal{D}_\alpha \to \mathcal{D}_\alpha \to \mathbb{B}$, the intended meaning of $\top_o$ is T, and the intended meaning of $\bot_o$ is F. Often, $=_\alpha$ is written as an infix operator.

The intended meaning of $\top_o$ is T and the intended meaning of $\bot_o$ is F.

The intended meaning $\wedge_\alpha : \mathcal{D}_\alpha \to \mathcal{D}_\alpha \to \mathcal{D}_\alpha$ of $\wedge_\alpha$ is defined by induction on the rank of α. For the base case, the intended meaning $\wedge_o : \mathbb{B} \to \mathbb{B} \to \mathbb{B}$ of $\wedge_o$ is $\wedge$, the usual conjunction connective. For the inductive step, assuming the intended meaning $\wedge_{\mathsf{t}(\alpha)} : \mathcal{D}_{\mathsf{t}(\alpha)} \to \mathcal{D}_{\mathsf{t}(\alpha)} \to \mathcal{D}_{\mathsf{t}(\alpha)}$ of $\wedge_{\mathsf{t}}(\alpha)$ has been defined, then the intended meaning $\wedge_\alpha$ of $\wedge_\alpha$ is

$$\lambda x.\lambda y.\lambda z_1.(x(z_1) \wedge_{\mathsf{t}(\alpha)} y(z_1)).$$

The intended meanings of the other connectives $\neg_\alpha$, $\vee_\alpha$, $\longrightarrow_\alpha$, and $\longleftarrow_\alpha$ are defined analogously. Note that $\wedge_\alpha$, $\vee_\alpha$, $\longrightarrow_\alpha$, and $\longleftarrow_\alpha$ will often be written as infix operators.

Let α be a biterm type of rank n. The intended meaning $\boldsymbol{\Pi}_\alpha : \mathcal{D}_\alpha \to \mathbb{B}$ of Π_α is defined by

$$\boldsymbol{\Pi}_\alpha(x) = \begin{cases} \mathsf{T} & \text{if } x = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \\ \mathsf{F} & \text{otherwise.} \end{cases}$$

The intended meaning $\boldsymbol{\Sigma}_\alpha : \mathcal{D}_\alpha \to \mathbb{B}$ of Σ_α is defined by

$$\boldsymbol{\Sigma}_\alpha(x) = \begin{cases} \mathsf{T} & \text{if } x \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \\ \mathsf{F} & \text{otherwise.} \end{cases}$$

B.1.2 Terms

The next task is to define the central concept of a term. For the modal part of the definition, assume there are necessity modal operators $\Box_i$, for $i = 1, \ldots, m$.

Definition B.1.7. A *term*, together with its type, is defined inductively as follows.

1. A variable in $\mathfrak{V}$ of type α is a term of type α.
2. A constant in $\mathfrak{C}$ having signature α is a term of type α.
3. If t is a term of type β and x a variable of type α, then $\lambda x.t$ is a term of type $\alpha \to \beta$.
4. If s is a term of type $\alpha \to \beta$ and t a term of type α, then $(s\ t)$ is a term of type β.
5. If $t_1, \ldots, t_n$ are terms of type $\alpha_1, \ldots, \alpha_n$, respectively, then $(t_1, \ldots, t_n)$ is a term of type $\alpha_1 \times \cdots \times \alpha_n$ (for $n \geq 0$).
6. If t is a term of type α and $i \in \{1, \ldots, m\}$, then $\Box_i t$ is a term of type α.

Notation. $\mathfrak{L}$ denotes the set of all terms obtained from an alphabet and is called the *language* given by the alphabet.

Example B.1.2. The data constructors for constructing lists whose items have type α are $[]_\alpha : \mathit{List}\ \alpha$ and $\sharp_\alpha : \alpha \to \mathit{List}\ \alpha \to \mathit{List}\ \alpha$, where $\sharp_\alpha$ is usually written infix. The term $[]_\alpha$ represents the empty list and the term $s\,\sharp_\alpha\,t$ represents the list with head s and tail t, where the items have type α. Thus $4\,\sharp_{\mathit{Int}}\,5\,\sharp_{\mathit{Int}}\,6\,\sharp_{\mathit{Int}}\,[]_{\mathit{Int}}$ represents the list $[4,5,6]$, and $A\,\sharp_T\,B\,\sharp_T\,C\,\sharp_T\,[]_T$ represents the list $[A,B,C]$, where A, B, and C are constants of some type T, say.

The notation $[]_\alpha$ and $\sharp_\alpha$, with the subscript α indicating the type of items, is cumbersome, but is needed if one wants to explicitly distinguish the various constants. When polymorphism is introduced in Section B.1.10, the subscripts will be done away with (at the expense of possible confusion about exactly which constant is intended).

Next the class of biterms, a class of terms that includes the class of formulas and allows many concepts to be extended from formulas to biterms, is introduced.

Definition B.1.8. A *formula* is a term of type o.

Definition B.1.9. A *predicate* is a term having a type of the form $\alpha \to o$, for some α.

Definition B.1.10. A *biterm of rank* n is a term whose type is a biterm type of rank n.

A *biterm* is a biterm of rank n, for some $n \geq 0$.

Biterms are usually denoted by Greek symbols φ, ψ, and so on.

A biterm of rank 0 is a term having type o, that is, a formula; a biterm of rank 1 is a term having type $\alpha_1 \to o$, for some α_1, that is, a predicate; a biterm of rank 2 is a term having type $\alpha_1 \to \alpha_2 \to o$, for some α_1 and α_2; and so on. If φ is a biterm of rank n, then $\lambda x.\varphi$ is a biterm of rank $n+1$. If φ is a biterm of rank n and $(\varphi\ t)$ is a term, then $(\varphi\ t)$ is a biterm of rank $n-1$.

The quantifiers can be defined for biterms. Let t be a biterm of type β and x a variable of type α. The universal quantifier $\forall_\alpha$ is introduced by the abbreviation

$$\forall_\alpha x.t \triangleq (\Pi_{\alpha\to\beta}\ \lambda x.t)$$

and the existential quantifier $\exists_\alpha$ is introduced by the abbreviation

$$\exists_\alpha x.t \triangleq (\Sigma_{\alpha\to\beta}\ \lambda x.t).$$

Note that $\forall_\alpha x.t$ and $\exists_\alpha x.t$ are formulas (even when t is a biterm that is not a formula). Later it will be shown that $\exists_\alpha x.t$ and $\neg\forall_\alpha x.(\neg_\beta\ t)$ have the same meaning. Similarly, $\forall_\alpha x.t$ and $\neg\exists_\alpha x.(\neg_\beta\ t)$ have the same meaning.

The possibility modalities that are dual to the necessity modalities are now introduced. Let t be a biterm of type α. Then the possibility modalities $\Diamond_i$ are introduced by the abbreviation

$$\Diamond_i t \triangleq (\neg_\alpha\ \Box_i(\neg_\alpha\ t)),$$

for $i = 1,\ldots,m$. Note that while $\Box_i$ can be applied to arbitrary terms, $\Diamond_i$ can only be applied to biterms because negation only makes sense for biterms.

Example B.1.3. Condition 6 of Definition B.1.7 admits some rather odd looking terms. For example, one can write $\Box_i 42$ or $\Box_i A$, where A is a data constructor, or $\Box_i f$, where f is a constant, or even $\Box_i(A,42)$. All these kinds of modal terms and many others will be useful in practice.

Notation. The type subscripts on equality, the connectives, and the quantifiers are somewhat intrusive. So, from now on, relying on the context to make the type clear, the subscripts will be discarded. Thus $=_\alpha$ (resp., $\top_o$, $\bot_o$, $\neg_\alpha$, $\wedge_\alpha$, $\vee_\alpha$, $\longrightarrow_\alpha$, $\longleftarrow_\alpha$, Π_α, Σ_α, $\forall_\alpha$, $\exists_\alpha$) will be denoted by $=$ ($\top$, $\bot$, $\neg$, $\wedge$, $\vee$, $\longrightarrow$, $\longleftarrow$, Π, Σ, $\forall$, $\exists$).

A term of the form $(\neg\ t)$ is usually abbreviated to $\neg t$. So, for example, $(\neg\ \Box_i(\neg\ t))$ would be shortened to $\neg\Box_i\neg t$.

Application of modalities is right associative. Thus $(\Box_{j_1}(\Box_{j_2}\cdots(\Box_{j_r}t)\cdots))$ can be written as $\Box_{j_1}\Box_{j_2}\cdots\Box_{j_r}t$.

$\forall x_1.\ \ldots\ \forall x_n.\varphi$ is an abbreviation of $(\Pi\ \lambda x_1.\ \ldots\ \lambda x_n.\varphi)$.

$\exists x_1.\ \ldots\ \exists x_n.\varphi$ is an abbreviation of $(\Sigma\ \lambda x_1.\ \ldots\ \lambda x_n.\varphi)$.

Definition B.1.11. A term is *rigid* if each constant in the term is rigid.

In the semantics below, rigid terms have the same meaning in each world.

In Definition B.1.7, a term of the form $\lambda x.t$ in Part 3 is an *abstraction*, a term of the form $(s\ t)$ in Part 4 is an *application*, a term of the form $(t_1, \ldots, t_n)$ in Part 5 is a *tuple*, and a term of the form $\Box_i t$ in Part 6 is a *box term*.

More precisely, the meaning of Definition B.1.7 is as follows. Let $\mathfrak{E}$ denote the set of all expressions obtained from the alphabet, where an expression is a finite sequence of symbols drawn from the set of constants $\mathfrak{C}$, the set of variables $\mathfrak{V}$, $\Box_i$, for $i = 1, \ldots, m$, and '(', ')', 'λ','.', and ','. Then $\mathfrak{L}$ is the intersection of all sets $\mathfrak{X} \subseteq \mathfrak{E} \times \mathfrak{S}$ satisfying the following conditions.

1. If x is a variable in $\mathfrak{V}$ of type α (as a variable), then $(x, \alpha) \in \mathfrak{X}$.
2. If C is a constant in $\mathfrak{C}$ having signature α, then $(C, \alpha) \in \mathfrak{X}$.
3. If $(t, \beta) \in \mathfrak{X}$ and x is a variable of type α, then $(\lambda x.t, \alpha \to \beta) \in \mathfrak{X}$.
4. If $(s, \alpha \to \beta) \in \mathfrak{X}$ and $(t, \alpha) \in \mathfrak{X}$, then $((s\ t), \beta) \in \mathfrak{X}$.
5. If $(t_1, \alpha_1), \ldots, (t_n, \alpha_n) \in \mathfrak{X}$, then $((t_1, \ldots, t_n), \alpha_1 \times \cdots \times \alpha_n) \in \mathfrak{X}$ (for $n \geq 0$).
6. If $(t, \alpha) \in \mathfrak{X}$ and $i \in \{1, \ldots, m\}$, then $(\Box_i t, \alpha) \in \mathfrak{X}$.

There is always at least one set satisfying these conditions, namely $\mathfrak{E} \times \mathfrak{S}$. Thus the intersection is well defined and it satisfies Conditions 1 to 6. Hence $\mathfrak{L}$ is the smallest set satisfying Conditions 1 to 6.

Having made the meaning of Definition B.1.7 precise, it will be convenient to follow the usual practice and refer to just the t in some (t, α) as being the term. Thus the type of the term is sometimes suppressed. Since the type of a term is uniquely determined by the term, as is shown later, nothing is lost by this practice. This convention is applied immediately in the next proposition.

Proposition B.1.1. *Let $t \in \mathfrak{L}$. Then exactly one of the following conditions holds.*

1. *$t \in \mathfrak{V}$.*
2. *$t \in \mathfrak{C}$.*
3. *t has the form $\lambda x.s$, where $s \in \mathfrak{L}$.*

4. *t has the form $(u\ v)$, where $u, v \in \mathfrak{L}$.*

5. *t has the form $(t_1, \ldots, t_n)$, where $t_1, \ldots, t_n \in \mathfrak{L}$.*

6. *t has the form $\Box_i s$, where s is a term and $i \in \{1, \ldots, m\}$.*

Proof. It is clear that at most one of the conditions holds. Now suppose t is a term that is neither a variable, nor a constant, nor has the form $\lambda x.s$, $(u\ v)$, $(t_1, \ldots, t_n)$ or $\Box_i s$. Then $\mathfrak{L} \setminus \{t\}$ satisfies Conditions 1 to 6 in the definition of a term, which contradicts the definition of $\mathfrak{L}$ as being the smallest set satisfying Conditions 1 to 6. Thus t is either a variable, a constant, or has the form $\lambda x.s$, $(u\ v)$, $(t_1, \ldots, t_n)$ or $\Box_i s$.

Suppose that t has the form $\lambda x.s$, but that s is not a term. Then the set of terms $\mathfrak{L} \setminus \{t\}$ satisfies Conditions 1 to 6 in the definition of a term, which contradicts the definition of $\mathfrak{L}$ as being the smallest set satisfying Conditions 1 to 6. Thus s is a term. The arguments for Parts 4, 5 and 6 are similar. □

Proposition B.1.2.

1. *An expression of the form $\lambda x.t$ is a term iff t is a term.*

2. *An expression of the form $(s\ t)$ is a term iff s and t are terms, and s has type of the form $\alpha \to \beta$ and t has type α.*

3. *An expression of the form $(t_1, \ldots, t_n)$ is a term iff $t_1, \ldots, t_n$ are terms.*

4. *An expression of the form $\Box_i t$ is a term iff t is a term.*

Proof. 1. If $\lambda x.t$ is a term, then t is a term by Part 3 of Proposition B.1.1. Conversely, if t is a term, then $\lambda x.t$ is a term, since $\mathfrak{L}$ satisfies Condition 3 of the definition of a term.

2. Suppose that $(s\ t)$ is a term. Then s and t are terms by Part 4 of Proposition B.1.1. If either s does not have type of the form $\alpha \to \beta$ or t does not have type α, then $\mathfrak{L} \setminus \{(s\ t)\}$ satisfies Conditions 1 to 6 in the definition of a term, which contradicts the definition of $\mathfrak{L}$ as being the smallest set satisfying Conditions 1 to 6. Conversely, if s and t are terms, and s has type of the form $\alpha \to \beta$ and t has type α, then $(s\ t)$ is a term, since $\mathfrak{L}$ satisfies Condition 4 of the definition of a term.

3. Suppose that $(t_1, \ldots, t_n)$ is a term. Then $t_1, \ldots, t_n$ are terms by Part 5 of Proposition B.1.1. Conversely, if $t_1, \ldots, t_n$ are terms, then $(t_1, \ldots, t_n)$ is a term, since $\mathfrak{L}$ satisfies Condition 5 of the definition of a term.

4. Suppose that $\Box_i t$ is a term. Then t is a term by Part 6 of Proposition B.1.1. Conversely, if t is a term, then $\Box_i t$ is a term, since $\mathfrak{L}$ satisfies Condition 6 of the definition of a term. □

Proposition B.1.3. *The type of each term is unique.*

Proof. Suppose that some term t has types α and β, where $\alpha \neq \beta$. If t is neither a variable nor a constant, by repeated use of Proposition B.1.2, t must contain a term s that is either a variable or a constant with at least two distinct types. If s is a variable, one of these types must be distinct from the type of the variable (as a variable). If t is a constant, one of these types must be distinct from the signature of the constant. In either case, let this type be γ. Consider now the set $\mathfrak{L}$ with all those terms removed that contain s with type

γ. This set is strictly smaller than $\mathfrak{L}$ and satisfies Conditions 1 to 6 of the definition of a term, which gives a contradiction. □

Notation. For each $\alpha \in \mathfrak{S}$, $\mathfrak{L}_\alpha$ denotes the set of all terms of type α. Thus $\mathfrak{L} = \bigcup_{\alpha\in\mathfrak{S}} \mathfrak{L}_\alpha$.

In a higher-order logic, one may identify sets and predicates – the actual identification is between a set and its indicator (that is, characteristic) function which is a predicate. Thus, if t is of type o, the abstraction $\lambda x.t$ may be written as $\{x \mid t\}$ if it is intended to emphasize that its intended meaning is a set. The notation $\{\}$ means $\{x \mid \bot_o\}$. The notation $t \in s$ means $(s\ t)$, where s has type $\alpha \to o$, for some α. Furthermore, notwithstanding the fact that sets are mathematically identified with predicates, it is sometimes convenient to maintain an informal distinction between sets (as 'collections of objects') and predicates. For this reason, the notation $\{\alpha\}$ is introduced as a synonym for the type $\alpha \to o$. The term $(s\ t)$ is often written as simply $s\ t$, using juxtaposition to denote application. Juxtaposition is left associative, so that $r\ s\ t$ means $((r\ s)\ t)$.

To prove properties of terms, one can employ the following *principle of induction on the structure of terms.*

Proposition B.1.4. *Let $\mathfrak{X}$ be a subset of $\mathfrak{L}$ satisfying the following conditions.*

1. $\mathfrak{V} \subseteq \mathfrak{X}$.
2. $\mathfrak{C} \subseteq \mathfrak{X}$.
3. *If $t \in \mathfrak{X}$ and $x \in \mathfrak{V}$, then $\lambda x.t \in \mathfrak{X}$.*
4. *If $s, t \in \mathfrak{X}$, s has type $\alpha \to \beta$ and t has type α, then $(s\ t) \in \mathfrak{X}$.*
5. *If $t_1, \ldots, t_n \in \mathfrak{X}$, then $(t_1, \ldots, t_n) \in \mathfrak{X}$.*
6. *If $t \in \mathfrak{X}$, then $\Box_i t \in \mathfrak{X}$.*

Then $\mathfrak{X} = \mathfrak{L}$.

Proof. Clearly $\mathfrak{X}$ satisfies Conditions 1 to 6 of Definition B.1.7. Thus, since $\mathfrak{L}$ is the intersection of all such sets, it follows immediately that $\mathfrak{L} \subseteq \mathfrak{X}$. Thus $\mathfrak{X} = \mathfrak{L}$. □

In later inductive proofs about terms, Proposition B.1.4 will be the basis of the induction argument, but the appropriate set $\mathfrak{X}$ will never explicitly stated – in all cases, this should be immediately clear.

Definition B.1.12. The *free variables* of a term are defined inductively as follows.

1. The variable x is free in x.
2. A constant contains no free variables.
3. A variable other than x is free in $\lambda x.t$ if the variable is free in t.
4. A variable is free in $(s\ t)$ if the variable is free in s or t.
5. A variable is free in $(t_1, \ldots, t_n)$ if the variable is free in t_j, for some $j \in \{1, \ldots, n\}$.

6. A variable is free in $\Box_i t$ if the variable is free in t.

Definition B.1.13. A term is *closed* if it contains no free variables; otherwise, it is *open*.

Definition B.1.14. Let φ be a biterm with free variables $x_1, \ldots, x_n$, where x_i has type α_i, for $i = 1, \ldots, n$, and where the leftmost free occurrences in φ of the variables are in the order $x_1, \ldots, x_n$. Then the *universal closure* of φ, denoted $\forall(\varphi)$, is the biterm $\forall x_1. \ldots \forall x_n.\varphi$.

Note that, provided φ has at least one free variable, $\forall(\varphi)$ is a formula, even when φ is not a formula. If φ has no free variables, then $\forall(\varphi)$ is just φ.

B.1.3 Occurrences

The concept of an occurrence of a subterm of a term will be needed.

Notation. Let $\mathbb{N}^*$ denote the set of all strings over the alphabet of positive integers, with ε denoting the empty string.

Definition B.1.15. The *occurrence set* of a term t, denoted $\mathcal{O}(t)$, is the set of strings in $\mathbb{N}^*$ defined inductively as follows.

1. If t is a variable, then $\mathcal{O}(t) = \{\varepsilon\}$.
2. If t is a constant, then $\mathcal{O}(t) = \{\varepsilon\}$.
3. If t has the form $\lambda x.s$, then $\mathcal{O}(t) = \{\varepsilon\} \cup \{1o \mid o \in \mathcal{O}(s)\}$.
4. If t has the form $(u\ v)$, then $\mathcal{O}(t) = \{\varepsilon\} \cup \{1o \mid o \in \mathcal{O}(u)\} \cup \{2o' \mid o' \in \mathcal{O}(v)\}$.
5. If t has the form $(t_1, \ldots, t_n)$, then $\mathcal{O}(t) = \{\varepsilon\} \cup \bigcup_{i=1}^{n} \{io_i \mid o_i \in \mathcal{O}(t_i)\}$.
6. If t has the form $\Box_i s$, then $\mathcal{O}(t) = \{\varepsilon\} \cup \{1o \mid o \in \mathcal{O}(s)\}$.

Each $o \in \mathcal{O}(t)$ is called an *occurrence* in t.

More precisely, $\mathcal{O}$ is a function $\mathcal{O} : \mathfrak{L} \to 2^{\mathbb{N}^*}$ from the set of terms into the powerset of the set of all strings of positive integers. The existence and uniqueness of $\mathcal{O}$ depends upon that fact that $\mathfrak{L}$ is well founded under the substring relation and hence Proposition B.4.3 applies: $\mathcal{O}$ is defined directly on the minimal elements (that is, constants and variables) and is uniquely determined by the rules in the definition for abstractions, applications, tuples, and box terms.

Definition B.1.16. If t is a term and $o \in \mathcal{O}(t)$, then the *subterm of t at occurrence o*, denoted $t|_o$, is defined inductively on the length of o as follows.

1. If $o = \varepsilon$, then $t|_o = t$.
2. If $o = 1o'$, for some o', and t has the form $\lambda x.s$, then $t|_o = s|_{o'}$.
   If $o = 1o'$, for some o', and t has the form $(u\ v)$, then $t|_o = u|_{o'}$.
   If $o = 2o'$, for some o', and t has the form $(u\ v)$, then $t|_o = v|_{o'}$.
   If $o = io'$, for some o', and t has the form $(t_1, \ldots, t_n)$, then $t|_o = t_i|_{o'}$, for $i = 1, \ldots, n$.
   If $o = 1o'$, for some o', and t has the form $\Box_i s$, then $t|_o = s|_{o'}$.

A *subterm* is a subterm of a term at some occurrence. A subterm is *proper* if it is not at occurrence ε.

Note that a variable appearing immediately after a λ in a term is *not* a subterm since the variable appearing there is not at an occurrence of the term.

An induction argument shows that each subterm is a term.

Definition B.1.17. An occurrence of a variable x in a term is *bound* if it occurs within a subterm of the form $\lambda x.t$.

A variable in a term is *bound* if it has a bound occurrence.

An occurrence of a variable in a term is *free* if it is not a bound occurrence.

For a particular occurrence of a subterm $\lambda x.t$ in a term, the occurrence of t is called the *scope* of the λx.

An induction argument shows that a variable is free in a term iff it has a free occurrence in the term.

Definition B.1.18. For a particular occurrence of a subterm $\Box_i t$ in a term, the occurrence of t is called the *scope* of the $\Box_i$.

An occurrence is *modal* if it occurs within the scope of a subterm of the form $\Box_i t$.

A related concept is that of the modal path to an occurrence in a term. Intuitively, this is the sequence of indexes of the modalities that are encountered on the path from the root of the term down to the occurrence.

Definition B.1.19. If t is a term and $o \in \mathcal{O}(t)$, then the *modal path to o in t* is defined inductively on the length of o as follows.

1. If $o = \varepsilon$, then the modal path is empty.
2. If $o = 1o'$, for some o', and t has the form $\lambda x.s$, then the modal path to o in t is the same as the modal path to o' in s.
   If $o = 1o'$, for some o', and t has the form $(u\ v)$, then the modal path to o in t is the same as the modal path to o' in u.
   If $o = 2o'$, for some o', and t has the form $(u\ v)$, then the modal path to o in t is the same as the modal path to o' in v.
   If $o = io'$, for some o', and t has the form $(t_1, \ldots, t_n)$, then the modal path to o in t is the same as the modal path to o' in t_i, for $i = 1, \ldots, n$.
   If $o = 1o'$, for some o', and t has the form $\Box_i s$ and the modal path to o' in s is p, then the modal path to o in t is ip.

B.1.4 Substitutions

Next substitutions are introduced.

Definition B.1.20. A *substitution* is a finite set of the form $\{x_1/t_1, \ldots, x_n/t_n\}$, where each x_i is a variable, each t_i is a term distinct from x_i and having the same type as x_i, and $x_1, \ldots, x_n$ are distinct.

Each element x_i/t_i is called a *binding*. The set $\{x_1, \ldots, x_n\}$ is called the *domain* of the substitution. The set of free variables in $\{t_1, \ldots, t_n\}$ is called the *range* of the substitution.

The empty substitution containing no bindings is denoted by $\{\}$.

Let θ be a substitution and t a term. Then $\theta|_t$ is the substitution obtained from θ by restricting the domain of θ to just the free variables appearing in t.

Intuitively, the concept of instantiating a term t by a substitution $\theta \triangleq \{x_1/t_1, \ldots, x_n/t_n\}$, is simple – each free occurrence of a variable x_i in t is replaced by t_i. But there is a technical complication in that there may be a free variable y, say, in some t_i that is 'captured' in this process because, after instantiation, it occurs in the scope of a subterm of the form $\lambda y.s$ and therefore becomes bound in $t\theta$. Free variable capture spoils the intended meaning of instantiation and hence it is necessary to avoid it. There are two approaches to this: one can disallow instantiation if free variable capture would occur or one can rename bound variables in the term t to avoid free variable capture altogether. The latter approach is adopted here.

Definition B.1.21. Let t be a term and $\theta \triangleq \{x_1/t_1, \ldots, x_n/t_n\}$ a substitution. The *instance* $t\theta$ of t by θ is defined as follows.

1. If t is a variable x_i, for some $i \in \{1, \ldots, n\}$, then $x_i\theta = t_i$.
   If t is a variable y distinct from all the x_i, then $y\theta = y$.

2. If t is a constant C, then $C\theta = C$.

3. (a) If t is an abstraction $\lambda x.s$ such that, for all $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ implies x is not free in t_i, then

   $$(\lambda x.s)\theta = \lambda x.(s\theta|_{\lambda x.s}).$$

   (b) If t is an abstraction $\lambda x.s$ such that, for some $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ and x is free in t_i, then

   $$(\lambda x.s)\theta = \lambda y.(s(\{x/y\} \cup \theta|_{\lambda x.s})).$$

   (Here y is chosen to be the first variable of the same type as x that does not appear in $\lambda x.s$ or $\theta|_{\lambda x.s}$.)

4. If t is an application $(u\ v)$, then $(u\ v)\theta = (u\theta\ v\theta)$.

5. If t is a tuple $(t_1, \ldots, t_n)$, then $(t_1, \ldots, t_n)\theta = (t_1\theta, \ldots, t_n\theta)$.

6. If t is a box term $\Box_i s$, then $(\Box_i s)\theta = \Box_i(s\theta)$.

More precisely, what is being defined in Definition B.1.21 is a function

$$T : \mathfrak{L} \times \Theta \to \mathfrak{L}$$

such that $T(t, \theta) = t\theta$, for all $t \in \mathfrak{L}$ and $\theta \in \Theta$, where $\mathfrak{L}$ is the set of all terms and Θ is the set of all substitutions. To establish the existence of T, the principle of inductive construction on well-founded sets given by Proposition B.4.4 is employed. The difficulty that has to be coped with is that, for example, for $(u\theta\ v\theta)$ to be well-defined, it is not only necessary that $u\theta$ and $v\theta$ be terms, but that they have appropriate types. To set up the application of Proposition B.4.4, consider first the substring relation $\prec$ on $\mathfrak{L}$ and extend

this to a relation $\prec_2$ on $\mathfrak{L} \times \Theta$ defined as follows: $(s, \theta) \prec_2 (t, \psi)$ if $s \prec t$. It is easy to see that $\prec_2$ is a well-founded order on $\mathfrak{L} \times \Theta$. The minimal elements of $\mathfrak{L} \times \Theta$ are tuples of the form (t, θ), where t is either a variable or a constant. Second, two partitions are defined: $\{\mathfrak{L}_\alpha \times \Theta\}_{\alpha \in \mathfrak{S}}$ is a partition of $\mathfrak{L} \times \Theta$ and $\{\mathfrak{L}_\alpha\}_{\alpha \in \mathfrak{S}}$ is a partition of $\mathfrak{L}$. Then it needs to be checked that the two conditions concerning consistency in Proposition B.4.4 are satisfied. First, each minimal element is consistent since each instance of a variable or constant is defined to be a term of the same type. Second, by considering the last four cases in Definition B.1.21, it is clear that the rule defining T has the property that if (s, ψ) is consistent, for each $(s, \psi) \prec_2 (t, \theta)$, then (t, θ) is consistent. Thus, by Proposition B.4.4, the function T exists, is unique, and satisfies the condition $T(\mathfrak{L}_\alpha \times \Theta) \subseteq \mathfrak{L}_\alpha$, for all $\alpha \in \mathfrak{S}$. The latter condition states exactly that if t is a term of type α and θ a substitution, then $t\theta$ is a term of type α.

Proposition B.1.5. *Let t be a term and $\theta \triangleq \{x_1/t_1, \ldots, x_n/t_n\}$ a substitution. Then a variable x is free in $t\theta$ iff x is free in t and distinct from all x_i or, for some $i \in \{1, \ldots, n\}$, x is free in t_i and x_i is free in t.*

Proof. The proof is by induction on the structure of t.

Let t be a variable x_i, for some $i \in \{1, \ldots, n\}$. Then $x_i\theta = t_i$, and the result follows. Let t be a variable y distinct from all the x_i. Then $y\theta = y$, and the result follows.

Let t be a constant C. Then $C\theta = C$, and the result follows.

Let t be an abstraction $\lambda x.s$ such that, for all $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ implies x is not free in t_i. Then $(\lambda x.s)\theta = \lambda x.(s\theta|_{\lambda x.s})$. Suppose that $\{x_{i_1}, \ldots, x_{i_k}\}$ is the set of x_i that are free in $\lambda x.s$. Then

z is free in $(\lambda x.s)\theta$
iff $z \neq x$ and z is free in $s\theta|_{\lambda x.s}$ $[(\lambda x.s)\theta = \lambda x.(s\theta|_{\lambda x.s})]$
iff $z \neq x$ and (z is free in s and distinct from $x_{i_1}, \ldots, x_{i_k}$ or,
for some $i \in \{i_1, \ldots, i_k\}$, z is free in t_i and x_i is free in s) [Induction hypothesis]
iff z is free in $\lambda x.s$ and distinct from all x_i or,
for some $i \in \{1, \ldots, n\}$, z is free in t_i and x_i is free in $\lambda x.s$
[for all $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ implies x is not free in t_i].

Let t be an abstraction $\lambda x.s$ such that, for some $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ and x is free in t_i. Then $(\lambda x.s)\theta = \lambda y.(s(\{x/y\} \cup \theta|_{\lambda x.s}))$. Suppose that $\{x_{i_1}, \ldots, x_{i_k}\}$ is the set of x_i that are free in $\lambda x.s$. Then

z is free in $(\lambda x.s)\theta$
iff $z \neq y$ and z is free in $s(\{x/y\} \cup \theta|_{\lambda x.s})$ $[(\lambda x.s)\theta = \lambda y.(s(\{x/y\} \cup \theta|_{\lambda x.s}))]$
iff $z \neq y$ and (z is free in s and distinct from x and $x_{i_1}, \ldots, x_{i_k}$ or,
for some $i \in \{i_1, \ldots, i_k\}$, z is free in t_i and x_i is free in s or $z = y$ and x is free in s)
[Induction hypothesis]
iff z is free in $\lambda x.s$ and distinct from all x_i or,
for some $i \in \{1, \ldots, n\}$, z is free in t_i and x_i is free in $\lambda x.s$.

Let t be an application $(u\ v)$. Thus $(u\ v)\theta = (u\theta\ v\theta)$. Then

x is free in $(u\ v)\theta$
iff x is free in $u\theta$ or x is free in $v\theta$ $\quad[(u\ v)\theta = (u\theta\ v\theta)]$
iff (x is free in u and distinct from all x_i or,
for some $i \in \{1, \ldots, n\}$, x is free in t_i and x_i is free in u) or
(x is free in v and distinct from all x_i or,
for some $i \in \{1, \ldots, n\}$, x is free in t_i and x_i is free in v)
[Induction hypothesis]
iff x is free in $(u\ v)$ and distinct from all x_i or,
for some $i \in \{1, \ldots, n\}$, x is free in t_i and x_i is free in $(u\ v)$.

If t is a tuple $(t_1, \ldots, t_n)$, then the proof is similar to the preceding part.
Let t have the form $\Box_j s$. Thus $(\Box_j s)\theta = \Box_j(s\theta)$. Then

x is free in $(\Box_j s)\theta$
iff x is free in $s\theta$ $\quad[(\Box_j s)\theta = \Box_j(s\theta)]$
iff x is free in s and distinct from all x_i or,
for some $i \in \{1, \ldots, n\}$, x is free in t_i and x_i is free in s
[Induction hypothesis]
iff x is free in $\Box_j s$ and distinct from all x_i or,
for some $i \in \{1, \ldots, n\}$, x is free in t_i and x_i is free in $\Box_j s$.

□

Proposition B.1.6. *Let t be a term and θ a substitution. Then $t\theta = t\theta|_t$.*

Proof. If t is a variable or a constant, the result is obvious.

Suppose t is an abstraction $\lambda x.s$ such that, for all $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ implies x is not free in t_i. Then $(\lambda x.s)\theta = \lambda x.(s\theta|_{\lambda x.s}) = (\lambda x.s)\theta|_{\lambda x.s}$.

Suppose t is an abstraction $\lambda x.s$ such that, for some $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ and x is free in t_i. Then $(\lambda x.s)\theta = \lambda y.(s(\{x/y\} \cup \theta|_{\lambda x.s})) = (\lambda x.s)\theta|_{\lambda x.s}$. (Note that, for both $(\lambda x.s)\theta$ and $(\lambda x.s)\theta|_{\lambda x.s}$, the same variable y is chosen.)

Suppose t is an application $(u\ v)$. Then

$$
\begin{aligned}
&(u\ v)\theta \\
&= (u\theta\ v\theta) && \text{[Part 4 of Definition B.1.21]} \\
&= (u\theta|_u\ \ v\theta|_v) && \text{[Induction hypothesis]} \\
&= (u\theta|_{(u\ v)}\ \ v\theta|_{(u\ v)}) && \text{[Induction hypothesis]} \\
&= (u\ v)\theta|_{(u\ v)} && \text{[Part 4 of Definition B.1.21].}
\end{aligned}
$$

If t is a tuple or has the form $\Box_i s$, the proof is similar to the preceding case. □

Proposition B.1.7. *Let t be a term and $\theta \triangleq \{x_1/t_1, \ldots, x_n/t_n\}$ a substitution such that, for $i = 1, \ldots, n$, x_i is not free in t. Then $t\theta = t$.*

Proof. By Proposition B.1.6, $t\theta = t\{\}$. An easy induction argument on the structure of t then shows that $t\{\} = t$. □

Proposition B.1.8. *Let $\lambda x.s$ be a term and $\theta \triangleq \{x_1/t_1, \ldots, x_n/t_n\}$ a substitution such that, for $i = 1, \ldots, n$, x_i is free in $\lambda x.s$. Then the following hold.*

1. *If, for all i, x is not free in t_i, then*

$$(\lambda x.s)\theta = \lambda x.(s\theta).$$

2. *If, for some i, x is free in t_i, then*

$$(\lambda y.s)\theta = \lambda y.(s(\{x/y\} \cup \theta)),$$

where y is the first variable of the same type as x that does not appear in $\lambda x.s$ or θ.

Proof. The result follows easily from Definition B.1.21, since $\theta|_{\lambda x.s} = \theta$. □

Propositions B.1.6 and B.1.8 can simplify proofs of results involving substitutions. Typically, one can show that without loss of generality a substitution can be restricted to the free variables appearing in some term, by using Proposition B.1.6. Then Proposition B.1.8 shows that for application of the substitution to an abstraction only the simpler cases in that proposition need be considered. For an illustration of this approach, see the proof of Proposition B.2.27 below.

B.1.5 Term Replacement

For the equational reasoning introduced in Appendix B.3, it will be necessary to replace subterms of terms by other terms.

Definition B.1.22. Let t be a term, s a subterm of t at occurrence o, and r a term. Then the expression obtained by *replacing s in t by r*, denoted $t[s/r]_o$, is defined by induction on the length of o as follows.

If the length of o is 0, then $t[s/r]_o = r$.

For the inductive step, suppose the length of o is $n + 1$ ($n \geq 0$). There are several cases to consider.

If $o = 1o'$, for some o', and t has the form $\lambda x.w$, then $(\lambda x.w)[s/r]_o = \lambda x.(w[s/r]_{o'})$.

If $o = 1o'$, for some o', and t has the form $(u\ v)$, then $(u\ v)[s/r]_o = (u[s/r]_{o'}\ v)$.

If $o = 2o'$, for some o', and t has the form $(u\ v)$, then $(u\ v)[s/r]_o = (u\ v[s/r]_{o'})$.

If $o = io'$, for some $i \in \{1, \ldots, n\}$ and o', and t has the form $(t_1, \ldots, t_n)$, then $(t_1, \ldots, t_n)[s/r]_o = (t_1, \ldots, t_i[s/r]_{o'}, \ldots, t_n)$.

If $o = 1o'$, for some o', and t has the form $\Box_i q$, then $(\Box_i q)[s/r]_o = \Box_i(q[s/r]_{o'})$.

Proposition B.1.9. *Let t be a term, s a subterm of t at occurrence o, and r a term. Suppose that s and r have the same type. Then $t[s/r]_o$ is a term of the same type as t.*

Proof. The proof is by induction on the length n of o.

Suppose first that $n = 0$. Thus t is s and $t[s/r]_o$ is r, so that $t[s/r]_o$ is a term of the same type as t.

Suppose next that the result holds for occurrences of length n and o has length $n+1$. Thus t has the form $\lambda x.w$, $(u\ v)$, $(t_1,\ldots,t_n)$ or $\Box_i s$.

Consider the case when t has the form $\lambda x.w$ and $o = 1o'$, for some o'. Then $(\lambda x.w)[s/r]_o = \lambda x.(w[s/r]_{o'})$. By the induction hypothesis, $w[s/r]_{o'}$ is a term of the same type as w. Hence $(\lambda x.w)[s/r]_o$ is a term of the same type as $\lambda x.w$.

The other cases are similar. □

Replacing a subterm s in a term t by a term r is quite different to applying a substitution: subterms are replaced, not just free variables; and it is common, even desirable, for free variables in the replacement term r to be captured in $t[s/r]_o$ after replacement. Usually, all free variables in r are free variables in s, so that there is no *new* free variable that could be captured, but there is no reason to insist on this here.

B.1.6 α-Conversion

Sometimes it is convenient to be able to rename bound variables. In the next definition, the relation $\succ_\alpha$ corresponding to α-conversion is introduced for this purpose.

Definition B.1.23. The rule of *α-conversion* is as follows: $\lambda x.t \succ_\alpha \lambda y.(t\{x/y\})$, if y is not free in t and x does not occur freely in t in a subterm of the form $\lambda y.s$.

In the preceding definition, the bound variable x in the abstraction is replaced by the bound variable y. The condition that y not be free in t ensures the meaning of $\lambda x.t$ is preserved under the renaming. The condition that x does not occur freely in t in a subterm of the form $\lambda y.s$ ensures that no further renaming will be necessary when the substitution $\{x/y\}$ is applied to t. Clearly, if y does not occur in $\lambda x.t$ at all, both these conditions are satisfied.

Definition B.1.24. The relation $\longrightarrow_\alpha$ is defined by $u \longrightarrow_\alpha u[s/t]_o$ if s is a subterm of u at occurrence o and $s \succ_\alpha t$.

Let $\xrightarrow{*}_\alpha$ be the reflexive, transitive closure of $\longrightarrow_\alpha$.

Let $\xleftrightarrow{*}_\alpha$ be the reflexive, symmetric, and transitive closure of $\longrightarrow_\alpha$.

Definition B.1.25. If $s \xleftrightarrow{*}_\alpha t$, then s and t are said to be *α-equivalent*.

Note that α-equivalent terms differ only in the names of (some of) their bound variables. Also α-equivalent terms must have the same type.

Proposition B.1.10. *Let t and t' be terms. Then the following hold.*

1. *If $t \succ_\alpha t'$, then $t' \succ_\alpha t$.*
2. *If $t \xleftrightarrow{*}_\alpha t'$, then $t \xrightarrow{*}_\alpha t'$.*

Proof. 1. Suppose that $\lambda x.r \succ_\alpha \lambda y.(r\{x/y\})$. Thus y is not free in r and x does not occur freely in r in a subterm of the form $\lambda y.s$. Note that x is not free in $r\{x/y\}$ and y does not occur freely in $r\{x/y\}$ in a subterm of the form $\lambda x.s$. Also $\lambda x.((r\{x/y\})\{y/x\}) = \lambda x.r$. Hence $\lambda y.(r\{x/y\}) \succ_\alpha \lambda x.r$.

2. It suffices to show that if $t \longrightarrow_\alpha t'$, then $t' \longrightarrow_\alpha t$. But this follows immediately from Part 1. □

B.1.7 Composition of Substitutions

The definition of the composition of two substitutions is now given. In preparation for this, three useful results involving α-equivalence and substitutions are proved.

Proposition B.1.11. *Let $\lambda x.s$ be a term and $\theta \triangleq \{x_1/t_1, \ldots, x_n/t_n\}$ a substitution such that, for some $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ and x is free in t_i. Then there exists a variable y such that $\lambda x.s \succ_\alpha \lambda y.(s\{x/y\})$, $(\lambda x.s)\theta = (\lambda y.(s\{x/y\}))\theta$ and, for all $i \in \{1, \ldots, n\}$, x_i is free in $\lambda y.(s\{x/y\})$ implies y is not free in t_i.*

Proof. Choose y to be the first variable of the same type as x that does not appear in $\lambda x.s$ or $\theta|_{\lambda x.s}$. Then y is not free in s and x does not occur freely in s in a subterm of the form $\lambda y.r$. Hence $\lambda x.s \succ_\alpha \lambda y.(s\{x/y\})$.

Clearly, for all $i \in \{1, \ldots, n\}$, x_i is free in $\lambda y.(s\{x/y\})$ implies y is not free in t_i. Thus $(\lambda x.s)\theta = \lambda y.(s(\{x/y\} \cup \theta|_{\lambda x.s})) = \lambda y.((s\{x/y\})\theta|_{\lambda y.(s\{x/y\})}) = (\lambda y.(s\{x/y\}))\theta$. □

Proposition B.1.12. *Let t and t' be terms that are α-equivalent and θ a substitution. Then $t\theta$ is α-equivalent to $t'\theta$.*

Proof. Suppose that θ is $\{x_1/t_1, \ldots, x_n/t_n\}$. The proof is by induction on the structure of t.

If t is a variable or a constant, then the result is obvious.

Suppose that t has the form $\lambda x.s$. Then t' has the form $\lambda y.r$, where $\lambda x.s \overset{*}{\longleftrightarrow}_\alpha \lambda y.r$. By Proposition B.1.11, it can be assumed without loss of generality that, for all $i \in \{1, \ldots, n\}$, x_i is free in $\lambda x.s$ implies x is not free in t_i, and x_i is free in $\lambda y.r$ implies y is not free in t_i.

There are two cases to consider.

(a) t' has the form $\lambda x.r$, where $s \overset{*}{\longleftrightarrow}_\alpha r$.

By the induction hypothesis, $s\theta|_{\lambda x.s} \overset{*}{\longleftrightarrow}_\alpha r\theta|_{\lambda x.s}$. Then

$$\begin{aligned}
& (\lambda x.s)\theta \\
= \ & \lambda x.(s\theta|_{\lambda x.s}) \\
\overset{*}{\longleftrightarrow}_\alpha \ & \lambda x.(r\theta|_{\lambda x.r}) \qquad \text{[Induction hypothesis, since } \theta|_{\lambda x.s} = \theta|_{\lambda x.r}\text{]} \\
= \ & (\lambda x.r)\theta.
\end{aligned}$$

(b) t' has the form $\lambda y.r$, where $x \neq y$.

Since $\lambda x.s \overset{*}{\longleftrightarrow}_\alpha \lambda y.r$, it follows that $\lambda x.s \overset{*}{\longleftrightarrow}_\alpha \lambda x.w \succ_\alpha \lambda y.(w\{x/y\}) \overset{*}{\longleftrightarrow}_\alpha \lambda y.r$, for some term w. Thus, by Part (a), it suffices to show that $(\lambda x.w)\theta \overset{*}{\longleftrightarrow}_\alpha (\lambda y.(w\{x/y\}))\theta$. For this, note that y is not free in $w\theta|_{\lambda x.w}$ and x does not occur freely in $w\theta|_{\lambda x.w}$ in a subterm of the form $\lambda y.v$. Hence

$$\begin{aligned}
& (\lambda x.w)\theta \\
= \ & \lambda x.(w\theta|_{\lambda x.w}) \\
\succ_\alpha \ & \lambda y.((w\theta|_{\lambda x.w})\{x/y\}) \\
= \ & \lambda y.((w\{x/y\})\theta|_{\lambda y.(w\{x/y\})}) \\
= \ & (\lambda y.(w\{x/y\}))\theta.
\end{aligned}$$

Suppose that t has the form $(t_1, \ldots, t_n)$. Let t' be a term such that $t \overset{*}{\longleftrightarrow}_\alpha t'$. Thus $t' = (t'_1, \ldots, t'_n)$, where $t_i \overset{*}{\longleftrightarrow}_\alpha t'_i$, for $i = 1, \ldots, n$. By the induction hypothesis, $t_i\theta \overset{*}{\longleftrightarrow}_\alpha t'_i\theta$, for $i = 1, \ldots, n$. Then

$$\begin{aligned}
& (t_1, \ldots, t_n)\theta \\
= \ & (t_1\theta, \ldots, t_n\theta) \\
\overset{*}{\longleftrightarrow}_\alpha \ & (t'_1\theta, \ldots, t'_n\theta) \qquad \text{[Induction hypothesis]} \\
= \ & (t'_1, \ldots, t'_n)\theta.
\end{aligned}$$

If t has the form $(u\ v)$ or $\Box_i s$, the proof is similar to the preceding case.

□

Proposition B.1.13. *Let θ and φ be substitutions. Then the following hold.*

1. *There are only finitely many variables $x_1, \ldots, x_m$ such that $(x_i\theta)\varphi \neq x_i$, for $i = 1, \ldots, m$.*
2. *$t\{x_1/(x_1\theta)\varphi, \ldots, x_m/(x_m\theta)\varphi\}$ is α-equivalent to $(t\theta)\varphi$, for all terms t.*

Proof. 1. For any variable x that does not occur in a binding of the form x/t in θ or φ, it is clear that $(x\theta)\varphi = x$.

2. Using Proposition B.1.12, it can be assumed without loss of generality that no bound variable in t appears in θ or φ. The proof then proceeds by induction on the structure of the term t.

If t is a variable or a constant, the result is obvious.

Suppose that t has the form $\lambda x.s$. Let $x_{i_1}, \ldots, x_{i_p}$ be those x_i that are free in $\lambda x.s$. Put $\theta' = \theta|_{\lambda x.s}$ and $\varphi' = \varphi|_{\lambda x.(s\theta')}$. Let $x_{j_1}, \ldots, x_{j_q}$ be those x_i such that $(x_i\theta')\varphi' \neq x_i$. Then

$$\begin{aligned}
& (\lambda x.s)\{x_1/(x_1\theta)\varphi, \ldots, x_m/(x_m\theta)\varphi\} \\
= \ & \lambda x.(s\{x_1/(x_1\theta)\varphi, \ldots, x_m/(x_m\theta)\varphi\}|_{\lambda x.s}) \\
= \ & \lambda x.(s\{x_{i_1}/(x_{i_1}\theta)\varphi, \ldots, x_{i_p}/(x_{i_p}\theta)\varphi\}) \\
= \ & \lambda x.(s\{x_{i_1}/(x_{i_1}\theta')\varphi', \ldots, x_{i_p}/(x_{i_p}\theta')\varphi'\}) \\
= \ & \lambda x.(s\{x_{j_1}/(x_{j_1}\theta')\varphi', \ldots, x_{j_q}/(x_{j_q}\theta')\varphi'\}) \\
\overset{*}{\longleftrightarrow}_\alpha \ & \lambda x.((s\theta')\varphi') \qquad \text{[Induction hypothesis]} \\
= \ & (\lambda x.(s\theta'))\varphi \\
= \ & ((\lambda x.s)\theta)\varphi.
\end{aligned}$$

If t has the form $(t_1, \ldots, t_n)$, then

$$\begin{aligned}
& (t_1, \ldots, t_n)\{x_1/(x_1\theta)\varphi, \ldots, x_m/(x_m\theta)\varphi\} \\
= \ & (t_1\{x_1/(x_1\theta)\varphi, \ldots, x_m/(x_m\theta)\varphi\}, \ldots, t_n\{x_1/(x_1\theta)\varphi, \ldots, x_m/(x_m\theta)\varphi\}) \\
\overset{*}{\longleftrightarrow}_\alpha \ & ((t_1\theta)\varphi), \ldots, (t_n\theta)\varphi) \qquad \text{[Induction hypothesis]} \\
= \ & (t_1\theta, \ldots, t_n\theta)\varphi \\
= \ & ((t_1, \ldots, t_n)\theta)\varphi.
\end{aligned}$$

If t has the form $(u\ v)$ or $\Box_i s$, the proof is similar to the preceding case. □

Here is an example to show that α-equivalence cannot be replaced by identity in Part 2 of Proposition B.1.13. This is because bound variables may need to be renamed when a substitution is applied.

Example B.1.4. Let t be $\lambda x.y$, θ be $\{y/x\}$, and φ be $\{x/z\}$. Then $t\theta = \lambda w.x$, say, where w is a fresh variable. Thus $(t\theta)\varphi = \lambda w.z$. Now x_1 is y and x_2 is x. Thus $t\{x_1/(x_1\theta)\varphi, x_2/(x_2\theta)\varphi\} = (\lambda x.y)\{y/z, x/z\} = \lambda x.z \neq (t\theta)\varphi$.

Now the definition of composition can be given.

Definition B.1.26. Let θ and φ be substitutions. Then the *composition* $\theta \circ \varphi$ of θ and φ is the substitution $\{x_1/(x_1\theta)\varphi, \ldots, x_m/(x_m\theta)\varphi\}$, where $x_1, \ldots, x_m$ are the finitely many variables such that $(x_i\theta)\varphi \neq x_i$, for $i = 1, \ldots, m$.

According to Proposition B.1.13, if α-equivalence rather than identity is sufficient for the particular situation (and it nearly always is), one can calculate $t(\theta \circ \varphi)$ by calculating $(t\theta)\varphi$ instead.

B.1.8 Matching

For the computation system introduced below, given terms s and t, there will be a need to determine whether or not there is a substitution θ such that $s\theta$ is (α-equivalent to) t. This motivates the next definition.

Definition B.1.27. Let s and t be terms of the same type. Then a substitution θ is a *matcher* of s to t if $s\theta$ is α-equivalent to t. In this case, s is said to be *matchable* to t.

It follows immediately from the definition and the fact that $\overset{*}{\longleftrightarrow}_\alpha$ is an equivalence relation that, if θ and θ' are matchers of s to t, then $s\theta \overset{*}{\longleftrightarrow}_\alpha s\theta'$.

Here is a useful alternative formulation of the definition of matcher.

Proposition B.1.14. *Let s and t be terms of the same type. Then a substitution θ is a matcher of s to t iff there exist a sequence of terms $s_0 = s, s_1, \ldots, s_n$ and an increasing subsequence $i_1, \ldots, i_m$ of the sequence $1, \ldots, n$ such that*

1. *there exists a substitution θ_{i_j} such that $s_{i_j-1}\theta_{i_j} = s_{i_j}$, for $j = 1, \ldots, m$,*
2. *$s_{i-1} \overset{*}{\longleftrightarrow}_\alpha s_i$, for $i \in \{1, \ldots, n\} \setminus \{i_1, \ldots, i_m\}$,*
3. *$s_n \overset{*}{\longleftrightarrow}_\alpha t$, and*
4. *$\theta = \theta_{i_1} \circ \cdots \circ \theta_{i_m}$.*

Proof. Suppose first that θ is a matcher of s to t. In this case, put $n = 1$ and $\theta_1 = \theta$. Then Conditions 1 to 4 are clearly satisfied.

Conversely, suppose that there exist a sequence of terms $s_0 = s, s_1, \ldots, s_n$ and an increasing subsequence $i_1, \ldots, i_m$ of the sequence $1, \ldots, n$ satisfying Conditions 1 to 4. By repeated application of Proposition B.1.12, it follows that $(\ldots(s\theta_{i_1})\ldots)\theta_{i_m} \overset{*}{\longleftrightarrow}_\alpha s_n$. By Proposition B.1.13, $s(\theta_{i_1} \circ \cdots \circ \theta_{i_m}) \overset{*}{\longleftrightarrow}_\alpha (\ldots(s_0\theta_{i_1})\ldots)\theta_{i_m}$. Thus $s(\theta_{i_1} \circ \cdots \circ \theta_{i_m}) \overset{*}{\longleftrightarrow}_\alpha t$, and so $\theta_{i_1} \circ \cdots \circ \theta_{i_m}$ is a matcher of s to t. □

The matching algorithm in Figure B.1 determines whether one term is matchable with another. Note that the inputs to this algorithm are two terms that have no free variables in common. It is usual to standardize apart before applying a unification algorithm so doing this for matching as well is not out of the ordinary. If this condition were to be dropped, the algorithm in Figure B.1 would need modification.

function $Match(s,t)$ **returns** matcher θ, if s is matchable to t
failure, otherwise;

inputs: s and t, terms of the same type with no free variables in common;

$\theta := \{\}$;
while $s \neq t$ **do**

$o :=$ occurrence of innermost subterm containing symbol at leftmost point of disagreement between s and t;

if $s|_o$ has form $\lambda x.v$ and $t|_o$ has form $\lambda y.w$ % $x \neq y$

then

$s := s[\lambda x.v/\lambda z.(v\{x/z\})]_o$; % z is a variable not in s or t

$t := t[\lambda y.w/\lambda z.(w\{y/z\})]_o$;

else if $s|_o$ is a free occurrence of a variable x and there is no free occurrence of x in s to the left of o and each free occurrence of a variable in $t|_o$ is a free occurrence in t

then

$\theta := \theta \circ \{x/t|_o\}$;

$s := s\{x/t|_o\}$;

else return failure;

return θ;

Figure B.1: Algorithm for finding a matching substitution

The correctness and termination of this algorithm is now established.

Proposition B.1.15. *Let s and t be terms of the same type with no free variables in common. If s is matchable to t, then the algorithm in Figure B.1 terminates and returns a matcher of s to t. Otherwise, the algorithm terminates and returns failure.*

Proof. The algorithm does terminate since all terms obtained from t by changing the names of bound variables have the same fixed length as t and, in each iteration of the loop, the point of disagreement moves strictly to the right.

Suppose first that s is not matchable to t. In this case, the algorithm cannot return successfully, since this would imply that s was matchable to t, by Proposition B.1.14. Thus, since the algorithm terminates, it must return failure.

Suppose now that s is matchable to t. Let φ be a matcher of s to t, so that $s\varphi$ is α-equivalent to t. Now, at each step of the algorithm, either a further change of a bound variable is made to s and t or else a further binding of the form $\{x/t|_o\}$ is applied to s. Let the sequence of terms obtained by the algorithm starting from s by either changing a bound variable or else by applying a binding at each step be $s_0(= s), s_1, s_2, \ldots$. Similarly, let the sequence of terms obtained starting from t by either changing a bound variable or else doing nothing be $t_0(= t), t_1, t_2, \ldots$. It will be shown by induction on the number n of steps so far of the algorithm that there exists a substitution π_n such that the domain of π_n is a subset of the free variables in s, the range of π_n is a subset of the free variables in t, and $s_n\pi_n \overset{*}{\longleftrightarrow}_\alpha t_n$, for $n = 0, 1, \ldots$.

For $n = 0$, put $\pi_0 = \varphi|_s$. Then $s_o\pi_o \overset{*}{\longleftrightarrow}_\alpha t_0$, since $s\varphi \overset{*}{\longleftrightarrow}_\alpha t$. Since s and t have no free variables in common, the domain of π_0 is the set of free variables in s and the range is the set of free variables in t.

Next suppose that the result holds for n so that the domain of π_n is a subset of the free variables in s, the range of π_n is a subset of the free variables in t, and $s_n\pi_n \overset{*}{\longleftrightarrow}_\alpha t_n$. For the $(n + 1)$th step, there are two cases to consider.

If the disagreement is such that the $(n + 1)$th step of the algorithm is a change of bound variables, put $\pi_{n+1} = \pi_n$. Then $s_{n+1} \overset{*}{\longleftrightarrow}_\alpha s_n$, so that $s_{n+1}\pi_{n+1} \overset{*}{\longleftrightarrow}_\alpha s_n\pi_{n+1}$, by Proposition B.1.12. Now $s_n\pi_{n+1} = s_n\pi_n$, $s_n\pi_n \overset{*}{\longleftrightarrow}_\alpha t_n$ (by the induction hypothesis), and $t_n \overset{*}{\longleftrightarrow}_\alpha t_{n+1}$. Hence $s_{n+1}\pi_{n+1} \overset{*}{\longleftrightarrow}_\alpha t_{n+1}$.

Suppose next that the disagreement does not involve a change of bound variables. Then the condition immediately following the **else if** must hold. That is, the disagreement must be at a free occurrence o of a variable x and there must be no free occurrence of x in s_n to the left of o and each free occurrence of a variable in $t_n|_o$ must be a free occurrence in t_n; otherwise, it would not be true that $s_n\pi_n \overset{*}{\longleftrightarrow}_\alpha t_n$. Thus π_n must contain a binding of the form $\{x/r\}$, where r is α-equivalent to $t_n|_o$. Put $\pi_{n+1} = \pi_n \setminus \{x/r\}$. Clearly the domain of π_{n+1} is a subset of the free variables in s and the range of π_{n+1} is a subset of the free variables in t. Thus $\pi_n = \{x/r\} \circ \pi_{n+1}$. Also

$$
\begin{aligned}
& s_{n+1}\pi_{n+1} \\
&= (s_n\{x/t_n|_o\})\pi_{n+1} \\
&\overset{*}{\longleftrightarrow}_\alpha (s_n\{x/r\})\pi_{n+1} && \text{[Proposition B.1.12, since } r \text{ is } \alpha\text{-equivalent to } t_n|_o] \\
&\overset{*}{\longleftrightarrow}_\alpha s_n(\{x/r\} \circ \pi_{n+1}) && \text{[Proposition B.1.13]} \\
&= s_n\pi_n \\
&\overset{*}{\longleftrightarrow}_\alpha t_n && \text{[Induction hypothesis]} \\
&= t_{n+1}.
\end{aligned}
$$

This completes the induction argument.

In effect, it has been shown that in the case that s is matchable to t, the algorithm cannot terminate with failure.

Now suppose that s is matchable to t and the algorithm terminates at the nth step with $s_n = t_n$. Let the sequence of bindings obtained by the algorithm be $\theta_{i_1}, \ldots, \theta_{i_m}$. Put $\theta = \theta_{i_1} \circ \cdots \circ \theta_{i_m}$. Then Conditions 1 to 4 of Proposition B.1.14 are satisfied, so that θ is a matcher of s to t. Finally, it is clear from the algorithm that the domain of θ is a subset of the free variables of s. □

Here are three examples to illustrate the matching algorithm.

Example B.1.5. Let s be $\lambda x.(f\ x\ (g\ y\ z))$ and t be $\lambda z.(f\ z\ (g\ A\ B))$, where f, g, A, and B are constants with suitable signatures. Then the successive steps of the algorithm are as follows.

0.	$\lambda \underset{\uparrow}{x}.(f\ x\ (g\ y\ z))$	$\lambda \underset{\uparrow}{z}.(f\ z\ (g\ A\ B))$	
1.	$\lambda w.(f\ w\ (g\ \underset{\uparrow}{y}\ z))$	$\lambda w.(f\ w\ (g\ \underset{\uparrow}{A}\ B))$	$\{y/A\}$
2.	$\lambda w.(f\ w\ (g\ A\ \underset{\uparrow}{z}))$	$\lambda w.(f\ w\ (g\ A\ \underset{\uparrow}{B}))$	$\{z/B\}$
3.	$\lambda w.(f\ w\ (g\ A\ B))$	$\lambda w.(f\ w\ (g\ A\ B))$	

(The arrows indicate the points of disagreement and the substitutions in the last column are the substitutions applied at that step in the algorithm.) Thus $\lambda x.(f\ x\ (g\ y\ z))$ is matchable to $\lambda z.(f\ z\ (g\ A\ B))$ with matcher $\{y/A\} \circ \{z/B\}$.

Example B.1.6. Let s be $(f\ x\ (g\ x))$ and t be $(f\ y\ (g\ A))$. Then the successive steps of the algorithm are as follows.

0.	$(f\ \underset{\uparrow}{x}\ (g\ x))$	$(f\ \underset{\uparrow}{y}\ (g\ A))$	$\{x/y\}$
1.	$(f\ y\ (g\ \underset{\uparrow}{y}))$	$(f\ y\ (g\ \underset{\uparrow}{A}))$	

Thus $(f\ x\ (g\ x))$ is not matchable to $(f\ y\ (g\ A))$, since there is a free occurrence of y in s to the left of the point of disagreement. Note that, in contrast, s and t are unifiable.

Example B.1.7. Let s be $\lambda x.(f\ x\ y\ z)$ and t be $\lambda x.(f\ x\ A\ (g\ x))$. Then the successive steps of the algorithm are as follows.

0.	$\lambda x.(f\ x\ \underset{\uparrow}{y}\ z)$	$\lambda x.(f\ x\ \underset{\uparrow}{A}\ (g\ x))$	$\{y/A\}$
1.	$\lambda x.(f\ x\ A\ \underset{\uparrow}{z})$	$\lambda x.(f\ x\ A\ \underset{\uparrow}{(g\ x)})$	

Thus $\lambda x.(f\ x\ y\ z)$ is not matchable to $\lambda x.(f\ x\ A\ (g\ x))$, since x has a free occurrence in $(g\ x)$ but this occurrence is not free in $\lambda x.(f\ x\ A\ (g\ x))$.

B.1.9 Representation of Individuals

In this section, the application of the logic to the representation of individuals is studied. The main idea is the identification of a class of terms, called *basic terms*, suitable for representing individuals in diverse applications. The most interesting aspect of the class of basic terms is that it includes certain abstractions and therefore is wider than is normally considered for knowledge representation. These abstractions allow one to model sets, multisets, and data of similar types, in a direct way. To define basic terms, one first needs to define normal terms, so the discussion starts there. This section is brief since this topic is covered in detail elsewhere. (See the Bibliographical Notes.)

First, some motivation. The term 'individual' is used to mean an item worth identifying and representing in an application; other terminology for the same idea includes 'object' and 'instance'. How should individuals be represented? Well there are many possible data types for this, including integers, natural numbers, floats, characters, strings, and booleans; tuples; sets; multisets; lists; trees; and graphs. Put this way, there does not appear to

be much structure in the way these data types are put together to represent individuals. However, it turns out that the data types above can be grouped into three major classes, constructor-based ones, abstractions, and tuples, and this grouping is the basis of the definition of basic terms. The first and third classes of data types are already widely used; for example, in functional programming languages. It is the second class which includes sets and multisets that is the most interesting of the three, so it is considered in more detail.

How should a (finite) set or multiset be represented? First, advantage is taken of the higher-order nature of the logic to identify sets and their indicator functions; that is, sets are viewed as predicates. With this approach, an obvious representation of sets uses the connectives, so that $\lambda x.(x = 1) \vee (x = 2)$ is the representation of the set $\{1, 2\}$. This kind of representation works well for sets. But the connectives are, of course, not available for multisets, so something more general is needed. An alternative representation for the set $\{1, 2\}$ is the term

$$\lambda x.\mathit{if}\ x = 1\ \mathit{then}\ \top\ \mathit{else}\ \mathit{if}\ x = 2\ \mathit{then}\ \top\ \mathit{else}\ \bot,$$

and this idea generalizes to multisets and similar abstractions. For example,

$$\lambda x.\mathit{if}\ x = A\ \mathit{then}\ 42\ \mathit{else}\ \mathit{if}\ x = B\ \mathit{then}\ 21\ \mathit{else}\ 0$$

is the multiset with 42 occurrences of A and 21 occurrences of B (and nothing else). Thus abstractions of the form

$$\lambda x.\mathit{if}\ x = t_1\ \mathit{then}\ s_1\ \mathit{else}\ \ldots\ \mathit{if}\ x = t_n\ \mathit{then}\ s_n\ \mathit{else}\ s_0$$

are adopted to represent (extensional) sets, multisets, and so on.

However, before giving the definition of a normal term, some attention has to be paid to the term s_0 in the previous expression. The reason is that s_0 in this abstraction is usually a very specific term. For example, for finite sets, s_0 is $\bot$ and for finite multisets, s_0 is 0. For this reason, the concept of a default term is now introduced. The intuitive idea is that, for each type, there is a (unique) default term such that each abstraction having that type as codomain takes the default term as its value for all but a finite number of points in the domain, that is, s_0 is the default value. The choice of default term depends on the particular application but, since sets and multisets are so useful, one would expect the set of default terms to include $\bot$ and 0. However, there could also be other types for which a default term is needed.

For each type constructor T and types $\alpha_1, \ldots, \alpha_k$, Assume there is chosen a unique *default data constructor* C such that C has signature of the form $\sigma_1 \to \cdots \to \sigma_n \to (T\ \alpha_1 \ldots \alpha_k)$. For example, for o, the default data constructor could be $\bot$, for *Int*, the default data constructor could be 0, and for *List* α, the default data constructor could be $[]_\alpha$, for all α.

Definition B.1.28. The set of *default terms*, $\mathfrak{D}$, is defined inductively as follows.

1. If C is a default data constructor of arity n and $t_1, \ldots, t_n \in \mathfrak{D}$ $(n \in \mathbb{N}_0)$ such that $C\ t_1 \ldots t_n \in \mathfrak{L}$, then $C\ t_1 \ldots t_n \in \mathfrak{D}$.

2. If $t \in \mathfrak{D}$ and $x \in \mathfrak{V}$, then $\lambda x.t \in \mathfrak{D}$.

3. If $t_1, \ldots, t_n \in \mathfrak{D}$ $(n \in \mathbb{N}_0)$, then $(t_1, \ldots, t_n) \in \mathfrak{D}$.

Definition B.1.29. For each $\alpha \in \mathfrak{S}$, define $\mathfrak{D}_\alpha = \{t \in \mathfrak{D} \mid t \text{ has type } \alpha\}$.

Now normal terms can be defined.

Definition B.1.30. The set of *normal terms*, $\mathfrak{N}$, is defined inductively as follows.

1. If C is a data constructor of arity n and $t_1, \ldots, t_n \in \mathfrak{N}$ $(n \in \mathbb{N}_0)$ such that $C\ t_1 \ldots t_n \in \mathfrak{L}$, then $C\ t_1 \ldots t_n \in \mathfrak{N}$.

2. If $t_1, \ldots, t_n \in \mathfrak{N}$, $s_1, \ldots, s_n \in \mathfrak{N}$ $(n \in \mathbb{N}_0)$, $s_0 \in \mathfrak{D}$ and

   $$\lambda x.\mathit{if}\ x = t_1\ \mathit{then}\ s_1\ \mathit{else}\ \ldots\ \mathit{if}\ x = t_n\ \mathit{then}\ s_n\ \mathit{else}\ s_0 \in \mathfrak{L},$$

   then

   $$\lambda x.\mathit{if}\ x = t_1\ \mathit{then}\ s_1\ \mathit{else}\ \ldots\ \mathit{if}\ x = t_n\ \mathit{then}\ s_n\ \mathit{else}\ s_0 \in \mathfrak{N}.$$

3. If $t_1, \ldots, t_n \in \mathfrak{N}$ $(n \in \mathbb{N}_0)$, then $(t_1, \ldots, t_n) \in \mathfrak{N}$.

Part 1 of the definition of the set of normal terms states, in particular, that individual natural numbers, integers, and so on, are normal terms. Also a term formed by applying a data constructor to (all of) its arguments, each of which is a normal term, is a normal term. As an example of this, consider the following declarations of the data constructors *Circle* and *Rectangle*.

$$\mathit{Circle} : \mathit{Float} \to \mathit{Shape}$$
$$\mathit{Rectangle} : \mathit{Float} \to \mathit{Float} \to \mathit{Shape}.$$

Then (*Circle* 7.5) and (*Rectangle* 42.0 21.3) are normal terms of type *Shape*. However, (*Rectangle* 42.0) is not a normal term as not all arguments to *Rectangle* are given. Normal terms coming from Part 1 of the definition are called *normal structures* and always have a type of the form $T\ \alpha_1 \ldots \alpha_k$.

The abstractions formed in Part 2 of the definition are 'almost constant' abstractions since they take the default term s_0 as value for all except a finite number of points in the domain. (The term s_0 is called the *default value* for the abstraction.) They are called *normal abstractions* and always have a type of the form $\beta \to \gamma$. This class of abstractions includes useful data types such as (finite) sets and multisets. More generally, normal abstractions can be regarded as lookup tables, with s_0 as the value for items not in the table.

Part 3 of the definition of normal terms just states that one can form a tuple from normal terms and obtain a normal term. These terms are called *normal tuples* and always have a type of the form $\alpha_1 \times \cdots \times \alpha_n$.

Normal terms are not quite what is wanted because, for example, they do not give a unique representation of sets. The problem is that the items of the set can appear in any order in a normal term representing the set. The obvious solution is to define a total order on normal terms and then use this order to give a unique representation of sets and other abstractions. Assuming that this total order $<$ is available, basic terms can now be defined.

Definition B.1.31. The set of *basic terms*, $\mathfrak{B}$, is defined inductively as follows.

1. If C is a data constructor of arity n and $t_1, \ldots, t_n \in \mathfrak{B}$ $(n \in \mathbb{N}_0)$ such that $C\ t_1 \ldots t_n \in \mathfrak{L}$, then $C\ t_1 \ldots t_n \in \mathfrak{B}$.

2. If $t_1, \ldots, t_n \in \mathfrak{B}$, $s_1, \ldots, s_n \in \mathfrak{B}$, $t_1 < \cdots < t_n$, $s_i \notin \mathfrak{D}$, for $1 \leq i \leq n$ $(n \in \mathbb{N}_0)$, $s_0 \in \mathfrak{D}$ and

   $$\lambda x.\mathit{if}\ x = t_1\ \mathit{then}\ s_1\ \mathit{else}\ \ldots\ \mathit{if}\ x = t_n\ \mathit{then}\ s_n\ \mathit{else}\ s_0 \in \mathfrak{L},$$

   then

   $$\lambda x.\mathit{if}\ x = t_1\ \mathit{then}\ s_1\ \mathit{else}\ \ldots\ \mathit{if}\ x = t_n\ \mathit{then}\ s_n\ \mathit{else}\ s_0 \in \mathfrak{B}.$$

3. If $t_1, \ldots, t_n \in \mathfrak{B}$ $(n \in \mathbb{N}_0)$, then $(t_1, \ldots, t_n) \in \mathfrak{B}$.

The basic terms from Part 1 of the definition are called *basic structures*, those from Part 2 are called *basic abstractions*, and those from Part 3 are called *basic tuples*.

Definition B.1.32. For each $\alpha \in \mathfrak{S}$, define $\mathfrak{B}_\alpha = \{t \in \mathfrak{B} \mid t \text{ has type } \alpha\}$.

The sets $\{\mathfrak{B}_\alpha\}_{\alpha \in \mathfrak{S}}$ play an important role in knowledge representation. For example, for a particular application, the representation space of each class of individuals would be $\mathfrak{B}_\alpha$, for some choice of $\alpha \in \mathfrak{S}$.

B.1.10 Polymorphism

In most applications, (parametric) polymorphism arises naturally. This section shows how polymorphism can be introduced into the logic.

Types To begin with, the earlier concept of a type has to be generalized. For this purpose, parameters, which are type variables, are introduced.

Definition B.1.33. An *alphabet* consists of four sets:

1. A set $\mathfrak{T}$ of type constructors.

2. A set $\mathfrak{P}$ of parameters.

3. A set $\mathfrak{C}$ of constants.

4. A set $\mathfrak{V}$ of variables.

Another difference in the above definition of an alphabet compared to the earlier one is that the variables in $\mathfrak{V}$ do not have fixed types; instead each occurrence of a variable in a term is given a type according to its position in the term.

Definition B.1.34. A *type* is defined inductively as follows.

1. Each parameter in $\mathfrak{P}$ is a type.

2. If T is a type constructor in $\mathfrak{T}$ of arity k and $\alpha_1, \ldots, \alpha_k$ are types, then $T\ \alpha_1 \ldots \alpha_k$ is a type. (For $k = 0$, this reduces to a type constructor of arity 0 being a type.)

3. If α and β are types, then $\alpha \to \beta$ is a type.

4. If $\alpha_1, \ldots, \alpha_n$ are types, then $\alpha_1 \times \cdots \times \alpha_n$ is a type.

Example B.1.8. Typical types are $Int \times (List\ a)$ and $List\ (List\ a)$.

Definition B.1.35. A type is *closed* if it contains no parameters.

A closed type in the sense of Definition B.1.35 is the same as a type in the sense of Definition B.1.2.

Since types may now contain parameters, the concept of a type substitution is introduced whose purpose is to instantiate parameters that appear in types.

Definition B.1.36. A *type substitution* is a finite set of the form $\{a_1/\alpha_1, \ldots, a_n/\alpha_n\}$, where each a_i is a parameter, each α_i is a type distinct from a_i, and $a_1, \ldots, a_n$ are distinct. Each element a_i/α_i is called a *binding*.

In particular, $\{\}$ is a type substitution called the *identity substitution*.

Notation. If $\mu = \{a_1/\alpha_1, \ldots, a_n/\alpha_n\}$, then $domain(\mu) = \{a_1, \ldots, a_n\}$ and $range(\mu)$ is the set of parameters appearing in $\{\alpha_1, \ldots, \alpha_n\}$.

Definition B.1.37. A type substitution $\{a_1/\alpha_1, \ldots, a_n/\alpha_n\}$ is *closed* if α_i is closed, for $i = 1, \ldots, n$.

Definition B.1.38. Let $\mu = \{a_1/\alpha_1, \ldots, a_n/\alpha_n\}$ be a type substitution and α a type. Then $\alpha\mu$, the *instance* of α by μ, is the type obtained from α by simultaneously replacing each occurrence of the parameter a_i in α by the type α_i $(i = 1, \ldots, n)$.

Definition B.1.39. Let $\mu = \{a_1/\alpha_1, \ldots, a_m/\alpha_m\}$ and $\nu = \{b_1/\beta_1, \ldots, b_n/\beta_n\}$ be type substitutions. Then the *composition* $\mu\nu$ of μ and ν is the type substitution obtained from the set

$$\{a_1/\alpha_1\nu, \ldots, a_m/\alpha_m\nu, b_1/\beta_1, \ldots, b_n/\beta_n\}$$

by deleting any binding $a_i/\alpha_i\nu$ for which $a_i = \alpha_i\nu$ and deleting any binding b_j/β_j for which $b_j \in \{a_1, \ldots, a_m\}$.

Composition is defined so that $\alpha(\mu\nu) = (\alpha\mu)\nu$, for any type α and type substitutions μ and ν. Also $\mu\{\} = \{\}\mu = \mu$, for any μ, so $\{\}$ really is an identity. Composition of type substitutions is associative.

One type can be more general than another.

Definition B.1.40. Let α and β be types. Then α is *more general than* β if there exists a type substitution ξ such that $\beta = \alpha\xi$.

Note that 'more general than' includes 'equal to', since ξ can be the identity substitution.

Example B.1.9. Let $\alpha = (List\ a) \times o$ and $\beta = (List\ Int) \times o$. Then α is more general than β, since $\beta = \alpha\xi$, where $\xi = \{a/Int\}$.

Definition B.1.41. Let $E = \{\alpha_1 = \beta_1, \ldots, \alpha_n = \beta_n\}$ be a set of equations about types. Then a type substitution μ is a *unifier* for E if $\alpha_i\mu$ is identical to $\beta_i\mu$, for $i = 1, \ldots, n$.

Example B.1.10. Let $E = \{a \to List\ b = c \to List\ M, a \times b = a \times M\}$, where a, b and c are parameters and M is a unary type constructor. Then the type substitution $\{a/M, b/M, c/M\}$ is a unifier for E.

One type substitution can be more general than another.

Definition B.1.42. Let μ and ν be type substitutions. Then μ is *more general* than ν if there exists a type substitution γ such that $\mu\gamma = \nu$.

Definition B.1.43. Let E be a set of equations about types and μ be a unifier for E. Then μ is a *most general unifier* for E if, for every unifier ν for E, μ is more general than ν.

Notation. The phrase 'most general unifier' is often abbreviated to 'mgu'.

Example B.1.11. Let $E = \{a \to List\ b = c \to List\ M, a \times b = a \times M\}$ be a set of equations about types. Then the type substitution $\{a/c, b/M\}$ is an mgu for E. Note that, for the unifier $\{a/M, b/M, c/M\}$ of Example B.1.10, $\{a/M, b/M, c/M\} = \{a/c, b/M\}\{c/M\}$. Thus $\{a/c, b/M\}$ is indeed more general than this unifier.

Terms Next the concept of a term from Section B.1.2 is generalized to the polymorphic case. For this, the discussion first turns to the constants in $\mathfrak{C}$.

In the polymorphic setting, constants can have signatures that are not closed, as is illustrated by the next example.

Example B.1.12. The data constructors for constructing lists are [] having signature $List\ a$ and $\sharp$ having signature $a \to List\ a \to List\ a$, where $\sharp$ is usually written infix. [] represents the empty list and the term $s \sharp t$ represents the list with head s and tail t. Because [] and $\sharp$ are polymorphic, [] and $s \sharp t$ can be used to represent lists of any type. Thus $4 \sharp 5 \sharp 6 \sharp []$ represents the list $[4, 5, 6]$, and $A \sharp B \sharp C \sharp []$ represents the list $[A, B, C]$, where A, B, and C are constants of some type T, say.

The set $\mathfrak{C}$ is similar to before, except that $=_\alpha$, $\neg_\alpha$, $\wedge_\alpha$, $\vee_\alpha$, $\longrightarrow_\alpha$, $\longleftarrow_\alpha$, Σ_α and Π_α are replaced by the following polymorphic versions (where a and the a_i are parameters):

1. $=$, having signature $a \to a \to o$.
2. $\neg$, having signature $(a_1 \to \cdots \to a_n \to o) \to (a_1 \to \cdots \to a_n \to o)$.
3. $\wedge$, $\vee$, $\longrightarrow$, and $\longleftarrow$, having signature $(a_1 \to \cdots \to a_n \to o) \to (a_1 \to \cdots \to a_n \to o) \to (a_1 \to \cdots \to a_n \to o)$.
4. Σ and Π, having signature $(a_1 \to \cdots \to a_n \to o) \to o$.

Data constructors always have a signature of the form $\sigma_1 \to \cdots \to \sigma_n \to (T\ a_1 \ldots a_k)$, where T is a type constructor of arity k, $a_1, \ldots, a_k$ are distinct parameters, and all the parameters appearing in $\sigma_1, \ldots, \sigma_n$ occur among $a_1, \ldots, a_k$ ($n \geq 0,\ k \geq 0$). The *arity* of the data constructor is n. A *nullary* data constructor is a data constructor of arity 0.

Now the concept of a (polymorphic) term, which generalizes that of a (monomorphic) term in Definition B.1.7, can be defined. However, the polymorphic case is rather more complicated than the monomorphic case since, when putting terms together to make larger terms, it is generally necessary to solve a system of equations and these equations depend upon the relative types of free variables in the component terms. The effect of this is that to define a term one has to define simultaneously its type, and its set of free variables and their relative types.

Definition B.1.44. A *term*, together with its type, and its set of free variables and their relative types, is defined inductively as follows.

1. Each variable x in $\mathfrak{V}$ is a term of type a, where a is a parameter.

   The variable x is free with relative type a in x.

2. Each constant C in $\mathfrak{C}$, where C has signature α, is a term of type α.

3. If t is a term of type β and x a variable in $\mathfrak{V}$, then $\lambda x.t$ is a term of type $\alpha \to \beta$, if x is free with relative type α in t, or type $a \to \beta$, where a is a new parameter, otherwise.

   A variable other than x is free with relative type σ in $\lambda x.t$ if the variable is free with relative type σ in t.

4. If s is a term of type α and t a term of type β such that the equation

   $$\alpha = \beta \to b,$$

   where b is a new parameter, augmented with equations of the form

   $$\varrho = \delta,$$

   for each variable that is free with relative type ϱ in s and is also free with relative type δ in t, have a most general unifier ξ, then $(s\ t)$ is a term of type $b\xi$.

   A variable is free with relative type $\sigma\xi$ in $(s\ t)$ if the variable is free with relative type σ in s or t.

5. If $t_1, \ldots, t_n$ are terms of type $\alpha_1, \ldots, \alpha_n$, respectively, such that the set of equations of the form

   $$\varrho_{i_1} = \cdots = \varrho_{i_k},$$

   for each variable that is free with relative type ϱ_{i_j} in the term t_{i_j} ($j = 1, \ldots, k$ and $k > 1$), have a most general unifier ξ, then $(t_1, \ldots, t_n)$ is a term of type $\alpha_1\xi \times \cdots \times \alpha_n\xi$.

   A variable is free with relative type $\sigma\xi$ in $(t_1, \ldots, t_n)$ if the variable is free with relative type σ in t_j, for some $j \in \{1, \ldots, n\}$.

6. If t is a term of type α and $i \in \{1, \ldots, m\}$, then $\Box_i t$ is a term of type α.

The type substitution ξ in Parts 4 and 5 of the definition is called the *associated* mgu.

Notation. Terms of the form $(\Sigma\ \lambda x.t)$ are written as $\exists x.t$ and terms of the form $(\Pi\ \lambda x.t)$ are written as $\forall x.t$.

Here are two examples to illustrate Definition B.1.44.

Example B.1.13. Let M be a nullary type, and $A : M$ and $concat : List\ a \times List\ a \to List\ a$ be constants. Recall that $[] : List\ a$ and $\sharp : a \to List\ a \to List\ a$ are the data constructors for lists. It will be shown that $(concat\ ([], [A]))$ is a term. For this, $([], [A])$ must be shown to be a term, which leads to the consideration of $[]$ and $[A]$. Now $[]$ is a term of type $List\ a$, by Part 2 of the definition of a term. By Parts 2 and 4, $(\sharp\ A)$ is a term of type $List\ M \to List\ M$, where along the way the equation $a = M$ is solved with the mgu $\{a/M\}$. Then $((\sharp\ A)\ [])$, which is the list $[A]$, is a term of type $List\ M$ by Part 4, where the equation $List\ M = List\ a$ is solved. By Part 5, it follows that $([], [A])$ is a term of type $List\ a \times List\ M$. Finally, by Part 4 again, $(concat\ ([], [A]))$ is a term of type $List\ M$, where the equation to be solved is $List\ a \times List\ a = List\ a \times List\ M$ whose mgu is $\{a/M\}$.

Example B.1.14. Consider the constants $append : List\ a \to List\ a \to List\ a \to o$ and $process : List\ a \to List\ a$. It will be shown that $\Box_i(((append\ x)\ [])\ (process\ x))$ is a term. First, the variable x is a term of type b, where the parameter is chosen to avoid a clash in the next step. Then $(append\ x)$ is a term of type $List\ a \to List\ a \to o$, for which the equation solved is $List\ a = b$. Next $((append\ x)\ [])$ is a term of type $List\ a \to o$ and x has relative type $List\ a$ in $((append\ x)\ [])$. Now consider $(process\ x)$, for which the constituent parts are $process$ of type $List\ c \to List\ c$ and the variable x of type d. Thus $(process\ x)$ is a term of type $List\ c$ and x has relative type $List\ c$ in $(process\ x)$. Finally, one has to apply $((append\ x)\ [])$ to the term $(process\ x)$. For this, by Part 4, there are two equations. These are $List\ a = List\ c$, coming from the top-level types, and $List\ a = List\ c$, coming from the free variable x in each of the components. These equations have the mgu $\{c/a\}$. Thus $(((append\ x)\ [])\ (process\ x))$ is a term of type o and hence, by Part 6, $\Box_i(((append\ x)\ [])\ (process\ x))$ is also a term of type o.

The concepts of an occurrence set in Definition B.1.15 and a subterm in Definition B.1.16 apply equally well to (polymorphic) terms given by Definition B.1.44. The concept of the relative type of a subterm in a (polymorphic) term will be useful.

Definition B.1.45. The *relative type* of the subterm $t|_o$ of the term t at $o \in \mathcal{O}(t)$ is defined by induction on the length of o as follows.

If the length of o is 0, then the relative type of $t|_o$ is the same as the type of t.

For the inductive step, suppose the length of o is $n + 1$ $(n \geq 0)$. There are several cases to consider.

If $o = 1o'$, for some o', and t has the form $\lambda x.s$, then the relative type of $t|_o$ is the same as the relative type of $s|_{o'}$ in s.

If $o = 1o'$, for some o', and t has the form $(u\ v)$, then the relative type of $t|_o$ is $\sigma\xi$, where σ is the relative type of $u|_{o'}$ in u and ξ is the associated mgu for $(u\ v)$.

If $o = 2o'$, for some o', and t has the form $(u\ v)$, then the relative type of $t|_o$ is $\sigma\xi$, where σ is the relative type of $v|_{o'}$ in v and ξ is the associated mgu for $(u\ v)$.

If $o = io'$, for some $i \in \{1, \ldots, n\}$ and o', and t has the form $(t_1, \ldots, t_n)$, then the relative type of $t|_o$ is $\sigma\xi$, where σ is the relative type of $t_i|_{o'}$ in t_i and ξ is the associated mgu for $(t_1, \ldots, t_n)$.

If $o = 1o'$, for some o', and t has the form $\Box_i s$, then the relative type of $t|_o$ is the same as the relative type of $s|_{o'}$ in s.

Example B.1.15. Let M be a nullary type, and $\mathit{append} : \mathit{List}\ a \to \mathit{List}\ a \to \mathit{List}\ a \to o$, $\mathit{process} : \mathit{List}\ a \to \mathit{List}\ a$ and $A, B, C : M$ be constants. Consider the first occurrence of x in the term $\Box_i(((\mathit{append}\ x)\ [])\ (\mathit{process}\ x))$. As a term in its own right, x has type a, for some parameter a. As a subterm of $\Box_i(((\mathit{append}\ x)\ [])\ (\mathit{process}\ x))$, x has relative type $\mathit{List}\ a$.

The occurrence of x in the term $\Box_i(((\mathit{append}\ x)\ [])\ (\mathit{process}\ [A, B, C]))$ has relative type $\mathit{List}\ M$. Also the subterm *process* of this term has relative type $\mathit{List}\ M \to \mathit{List}\ M$.

Next the exact connection between the two definitions of the concept of a term is established. Informally, a polymorphic term can be regarded as standing for a collection of monomorphic terms. A precise meaning to this last statement is now given.

For this purpose, attention must be paid to the alphabets that each definition of a term uses. In particular, given the alphabet that is used in Definition B.1.44, a suitable alphabet must be defined for use in Definition B.1.7. In both alphabets, the same set of type constructors is used, but the other components are different. Note that, for some fixed set of type constructors, the closed types in the sense of Definition B.1.34 are exactly the types in the sense of Definition B.1.2.

Definition B.1.46. Let A be an alphabet in the sense of Definition B.1.33 consisting of the set $\mathfrak{T}$ of type constructors, the set $\mathfrak{P}$ of parameters, the set $\mathfrak{C}$ of constants, and the set $\mathfrak{V}$ of variables. Then $\overline{A}$, the *underlying monomorphic alphabet* for A, is an alphabet in the sense of Definition B.1.1 given by the set $\mathfrak{T}$ of type constructors, the set $\overline{\mathfrak{C}}$ of constants, and the set $\overline{\mathfrak{V}}$ of variables, where

$$\overline{\mathfrak{C}} \triangleq \{C_\beta \mid C \in \mathfrak{C}, C \text{ has signature } \alpha, \text{ and } \beta \text{ is a closed instance of } \alpha\},$$

and

$$\overline{\mathfrak{V}} \triangleq \{x_\alpha \mid x \in \mathfrak{V} \text{ and } \alpha \text{ is a closed type}\}.$$

The preceding definition captures the intuition that a declaration for a polymorphic constant stands for the collection of declarations for the (monomorphic) constants that can be obtained by instantiating all parameters in the polymorphic declaration with closed types. The constant C_β in $\overline{\mathfrak{C}}$ has signature β (by definition). If $C \in \mathfrak{C}$ has closed signature α, then $C_\alpha \in \overline{\mathfrak{C}}$ also has signature α, and this is the only (monomorphic) constant that underlies C. By definition, a variable $x_\alpha \in \overline{\mathfrak{V}}$ has type α. The set $\overline{\mathfrak{V}}$ thus contains an infinite set of variables of each type.

Definition B.1.47. Let t be a term. A *grounding* type substitution for t is closed type substitution whose domain includes the set of all parameters in the relative types of subterms of t.

Definition B.1.48. Let A be an alphabet in the sense of Definition B.1.33, t be a term using A in the sense of Definition B.1.44, and γ a grounding type substitution for t. Then the expression t^γ using symbols from the alphabet $\overline{A}$ is defined inductively as follows:

1. If t is a variable x of type a, then x^γ is the variable $x_{a\gamma}$ in the alphabet $\overline{A}$.
2. If t is a constant constant C having signature α, then C^γ is the constant $C_{\alpha\gamma}$ in the alphabet $\overline{A}$.
3. If t is an abstraction $\lambda x.s$ and x is not free in s, then $(\lambda x.s)^\gamma$ is $\lambda x^\gamma.s^\gamma$; if x is free with relative type α in s, then $(\lambda x.s)^\gamma$ is $\lambda x^{\gamma'}.s^\gamma$, where x has type a and $\gamma' = \{a/\alpha\}\gamma$.
4. If t is an application $(u\ v)$, then $(u\ v)^\gamma$ is $(u^{\xi\gamma}\ v^{\xi\gamma})$, where ξ is the associated mgu.
5. If t is a tuple $(t_1, \ldots, t_n)$, then $(t_1, \ldots, t_n)^\gamma$ is $(t_1^{\xi\gamma}, \ldots, t_n^{\xi\gamma})$, where ξ is the associated mgu.
6. If t is a box term $\Box_i s$, then $(\Box_i s)^\gamma$ is $\Box_i s^\gamma$.

The next result shows that t^γ is a term in the sense of Definition B.1.7.

Proposition B.1.16. *Let A be an alphabet in the sense of Definition B.1.33, t a term of type α in the sense of Definition B.1.44 using A, and γ a grounding type substitution for t. Then t^γ is a term of type $\alpha\gamma$ in the sense of Definition B.1.1 using the alphabet $\overline{A}$.*

Proof. The proof uses the principle of inductive construction given in Proposition B.4.4. To apply this result, it is necessary to restate Definition B.1.48 more formally as the definition of a certain function. To set this up, let $\mathfrak{L}$ be the set of terms given by Definition B.1.44 for the alphabet A and $\overline{\mathfrak{L}}$ the set of terms given by Definition B.1.7 for the alphabet $\overline{A}$. Denote by $\mathfrak{G}$ the set of type substitutions. Let $\prec$ be the substring relation on $\mathfrak{L}$ and extend this to a relation $\prec_2$ on $\mathcal{L} \times \mathfrak{G}$ by $(s, \theta) \prec_2 (t, \psi)$ if $s \prec t$. Clearly $\prec_2$ is a well-founded order on $\mathfrak{L} \times \mathfrak{G}$

Put

$$\mathfrak{LG} = \{(t, \gamma) \mid t \in \mathfrak{L} \text{ and } \gamma \text{ is a grounding type substitution for } t\}.$$

Since $\mathfrak{LG}$ is a subset of $\mathfrak{L} \times \mathfrak{G}$, $\prec_2$ is a well-founded order on $\mathfrak{LG}$. The minimal elements in $\mathfrak{LG}$ are tuples (t, θ), where t is a variable or a constant.

Now define a function

$$U : \mathfrak{LG} \to \overline{\mathcal{L}}.$$

For this, two partitions are needed. Let $\mathfrak{S}$ be the set of closed types. For each $\alpha \in \mathfrak{S}$, let

$$\mathfrak{LG}_\alpha = \{(t, \gamma) \mid t \in \mathfrak{L}_\beta, \gamma \text{ is a grounding type substitution for } t \text{ and } \alpha = \beta\gamma\}.$$

Also, for each $\alpha \in \mathfrak{S}$, let $\overline{\mathfrak{L}}_\alpha$ be the set of terms in $\overline{\mathfrak{L}}$ of type α. Then $\{\mathfrak{LG}_\alpha\}_{\alpha\in\mathfrak{S}}$ is a partition of $\mathfrak{LG}$ and $\{\overline{\mathfrak{L}}_\alpha\}_{\alpha\in\mathfrak{S}}$ is a partition of $\overline{\mathfrak{L}}$. Now, using Proposition B.4.4, $U(t, \gamma)$ is, in effect, defined to be t^γ, for each $(t, \gamma) \in \mathfrak{LG}$.

First, consider the minimal elements in $\mathfrak{LG}$. If t is a variable x of type a, then define $U(x, \gamma)$ to be the variable $x_{a\gamma}$ in the alphabet $\overline{A}$. Also the consistency condition is satisfied

since $(x,\gamma) \in \mathfrak{LG}_{a\gamma}$ implies that $U(x,\gamma) \in \overline{\mathfrak{L}}_{a\gamma}$. If t is a constant C having signature α, then define $U(C,\gamma)$ to be the constant $C_{\alpha\gamma}$ in the alphabet $\overline{A}$. Once again the consistency is satisfied since $(x,\gamma) \in \mathfrak{LG}_{a\gamma}$ implies that $U(C,\gamma) \in \overline{\mathfrak{L}}_{a\gamma}$.

Next consider the non-minimal elements in $\mathfrak{LG}$. For these, the first component of the pair is an abstraction, application, tuple or box term.

Let t be an abstraction $\lambda x.s$, where x is not free in s and s has type β. Thus $\lambda x.s$ has type $a \to \beta$, where x has type a. Now $(s,\gamma) \in \mathfrak{LG}_{\beta\gamma}$ and $(s,\gamma) \prec_2 (\lambda x.s,\gamma)$. Assume that (s,γ) is consistent so that $U(s,\gamma) \in \overline{\mathfrak{L}}_{\beta\gamma}$. Now define $U(\lambda x.s,\gamma)$ to be $\lambda x^{\gamma}.s^{\gamma}$. This uniquely defines U on $(\lambda x.s,\gamma)$ and, furthermore, $(\lambda x.s,\gamma)$ is consistent since $(\lambda x.s,\gamma) \in \mathfrak{LG}_{(a\to\beta)\gamma}$ and $U(\lambda x.s,\gamma) \in \overline{\mathfrak{L}}_{(a\to\beta)\gamma}$.

Let t be an abstraction $\lambda x.s$, where x is free with relative type α in s and s has type β. Thus $\lambda x.s$ has type $\alpha \to \beta$. Now $(s,\gamma) \in \mathfrak{LG}_{\beta\gamma}$ and $(s,\gamma) \prec_2 (\lambda x.s,\gamma)$. Assume that (s,γ) is consistent so that $U(s,\gamma) \in \overline{\mathfrak{L}}_{\beta\gamma}$. Now define $U(\lambda x.s,\gamma)$ to be $\lambda x^{\gamma'}.s^{\gamma}$, where x has type a and $\gamma' = \{a/\alpha\}\gamma$. This uniquely defines U on $(\lambda x.s,\gamma)$ and, furthermore, $(\lambda x.s,\gamma)$ is consistent since $(\lambda x.s,\gamma) \in \mathfrak{LG}_{(\alpha\to\beta)\gamma}$ and $U(\lambda x.s,\gamma) \in \overline{\mathfrak{L}}_{(\alpha\to\beta)\gamma}$.

Let t be an application $(u\ v)$, where u has type α, v has type β, and ξ is the associated mgu. Thus ξ is an mgu of the equation $\alpha = \beta \to b$ and any additional equations that come from free variables common to u and v, and the type of $(u\ v)$ is $b\xi$. Now $(u,\xi\gamma) \in \mathfrak{LG}_{\alpha\xi\gamma}$ and $(u,\xi\gamma) \prec_2 ((u\ v),\gamma)$. Similarly, $(v,\xi\gamma) \in \mathfrak{LG}_{\beta\xi\gamma}$ and $(v,\xi\gamma) \prec_2 ((u\ v),\gamma)$. Assume that $(u,\xi\gamma)$ is consistent, so that $U(u,\xi\gamma) \in \overline{\mathfrak{L}}_{\alpha\xi\gamma}$ and that $(v,\xi\gamma)$ is consistent, so that $U(v,\xi\gamma) \in \overline{\mathfrak{L}}_{\beta\xi\gamma}$. Now define $U((u\ v),\gamma)$ to be $(u^{\xi\gamma}\ v^{\xi\gamma})$. This uniquely defines U on $((u\ v),\gamma)$ and, furthermore, $((u\ v),\gamma)$ is consistent since $((u\ v),\gamma) \in \mathfrak{LG}_{b\xi\gamma}$ and $U((u\ v),\gamma) \in \overline{\mathfrak{L}}_{b\xi\gamma}$.

Let t be a tuple $(t_1,\ldots,t_n)$, where t_i has type α_i, for $i = 1,\ldots,n$, and ξ the associated mgu. Thus the type of $(t_1,\ldots,t_n)$ is $\alpha_1\xi \times \cdots \times \alpha_n\xi$. Now $(t_i,\xi\gamma) \in \mathfrak{LG}_{\alpha_i\xi\gamma}$ and $(t_i,\xi\gamma) \prec_2 ((t_1,\ldots,t_n),\gamma)$, for $i = 1,\ldots,n$. Assume that $(t_i,\xi\gamma)$ is consistent, so that $U(t_i,\xi\gamma) \in \overline{\mathfrak{L}}_{\alpha_i\xi\gamma}$, for $i = 1,\ldots,n$. Now define $U((t_1,\ldots,t_n),\gamma)$ to be $(t_1^{\xi\gamma},\ldots,t_n^{\xi\gamma})$. This uniquely defines U on $((t_1,\ldots,t_n),\gamma)$ and, furthermore, $((t_1,\ldots,t_n),\gamma)$ is consistent since $((t_1,\ldots,t_n),\gamma) \in \mathfrak{LG}_{(\alpha_1\xi\times\cdots\times\alpha_n\xi)\gamma}$ and $U((t_1,\ldots,t_n),\gamma) \in \overline{\mathfrak{L}}_{(\alpha_1\xi\times\cdots\times\alpha_n\xi)\gamma}$.

Let t be a box term $\Box_i s$, for some $i \in \{1,\ldots,m\}$, where s has type α. Now $(s,\gamma) \in \mathfrak{LG}_{\alpha\gamma}$ and $(s,\gamma) \prec_2 (\Box_i s,\gamma)$. Assume that (s,γ) is consistent, so that $U(s,\gamma) \in \overline{\mathfrak{L}}_{\alpha\gamma}$. Now define $U(\Box_i s,\gamma)$ to be $\Box_i s^{\gamma}$. This uniquely defines U on $(\Box_i s,\gamma)$ and, furthermore, $(\Box_i s,\gamma)$ is consistent since $(\Box_i s,\gamma) \in \mathfrak{LG}_{\alpha\gamma}$ and $U(\Box_i s,\gamma) \in \overline{\mathfrak{L}}_{\alpha\gamma}$.

Proposition B.4.4 now shows that U is uniquely defined. Furthermore, $U(\mathfrak{LG}_{\alpha}) \subseteq \overline{\mathfrak{L}}_{\alpha}$, for each $\alpha \in \mathfrak{S}$. In other words, if t a term of type α in the sense of Definition B.1.44 using A and γ a grounding type substitution for t, then t^{γ} is a term of type $\alpha\gamma$ in the sense of Definition B.1.1 using the alphabet $\overline{A}$. □

Definition B.1.49. Let t be a term in the sense of Definition B.1.44 and γ a grounding type substitution for t. Then t^{γ} is called the *underlying monomorphic term* of t with respect to γ.

It is instructive to study the properties of the function U in Proposition B.1.16 on subterms.

Proposition B.1.17. *Let t be a term in the sense of Definition B.1.44 and γ a grounding type substitution for t. If o is an occurrence in t, then o is also an occurrence in t^{γ}.*

Furthermore, if s is a subterm of relative type σ at occurrence o in t, then the subterm of t^γ at occurrence o has type $\sigma\gamma$.

Proof. The first part is proved by induction on the length of the occurrence o in t. If the length of o is 0, then o is ε which is also an occurrence in t^γ. Suppose next that the result holds for occurrences of length n and o is an occurrence in t of length $n+1$. Thus t must be an abstraction, application, tuple or modal term. Suppose that t is a tuple $(t_1, \ldots, t_n)$ and $o = io'$, for some $i \in \{1, \ldots, n\}$. Now $(t_1, \ldots, t_n)^\gamma = (t_1^{\xi\gamma}, \ldots, t_n^{\xi\gamma})$, where ξ is the associated mgu. By the induction hypothesis, o' is an occurrence in $t_i^{\xi\gamma}$. Hence o is an occurrence in $(t_1, \ldots, t_n)^\gamma$. The other cases are similar.

The second part is also proved by induction on the length of the occurrence o in t. If the length of o is 0, then s is t and the subterm at occurrence o in t^γ, which has type $\alpha\gamma$, where α is the type of t, by Proposition B.1.16. Suppose next that the result holds for occurrences of length n and o is an occurrence in t of length $n+1$. Thus t must be an abstraction, application, tuple or modal term. Suppose that t is a tuple $(t_1, \ldots, t_n)$ and $o = io'$, for some $i \in \{1, \ldots, n\}$. Then s is a subterm at occurrence o' of t_i of relative type σ', where $\sigma = \sigma'\xi$. Now $(t_1, \ldots, t_n)^\gamma = (t_1^{\xi\gamma}, \ldots, t_n^{\xi\gamma})$, where ξ is the associated mgu. By the induction hypothesis, the subterm of $t_i^{\xi\gamma}$ at occurrence o' has type $\sigma'\xi\gamma = \sigma\gamma$. That is, the subterm at occurrence o in t^γ has type $\sigma\gamma$. The other cases are similar. □

In the following, when needed, signatures of polymorphic data constructors will be declared and definitions of polymorphic constants given. So, for example, the polymorphic constant *append* having signature

$$append \ : \ List\ a \times List\ a \times List\ a \to o$$

will be useful because appending lists is independent of the type of the items in the list. This declaration is thought of as a way of introducing the set of (monomorphic) constants of the form $append_{List\ \alpha \times List\ \alpha \times List\ \alpha \to o}$, where α ranges over all closed types. Generally, the type subscript is suppressed and the actual 'instance' of *append* intended is determined by the context. Similarly, one can give a polymorphic definition of the constant *append* as follows.

$$\begin{array}{l}(append\ (u, v, w)) = \\ \quad ((u = []) \wedge (v = w)) \vee \exists r.\exists x.\exists y.((u = r \sharp x) \wedge (w = r \sharp y) \wedge (append\ (x, v, y)))\end{array}$$

This is thought of as providing a definition for each of the underlying monomorphic constants by applying the appropriate grounding type substitution to the above definition.

The polymorphic syntax is extremely convenient and will be heavily exploited in the subsequent development because many common data constructors and constants are (parametrically) polymorphic. Furthermore, (non-nullary) type constructors are only interesting in the polymorphic case when they can take parameters as arguments. However, there is a cost associated with the use of polymorphism which is the need to infer the types of the variables and polymorphic constants from the context. So, for example, the monomorphic constant underlying *append* that appears in

$$(append\ (1 \sharp 2 \sharp [], 3 \sharp [], x))$$

can be inferred to be $append_{List\ Int \times List\ Int \times List\ Int \to o}$, since 1, 2, and 3 have type *Int*.

Generally, just the name itself of a polymorphic symbol (that is, constant or variable) will be used in a term, relying on the context to make clear which underlying monomorphic symbol is intended. For some purposes, it may be helpful to make explicit the grounding type substitution that is intended to be applied. For example, the term $(append\ (x, y, z))$ on its own does not determine the type of the monomorphic constant underlying *append*; this would then have to be given separately with a suitable grounding type substitution. An extreme case like this is that of a variable x of type a on its own that would need a grounding type substitution of the form $\{a/\alpha\}$ to specify the intended underlying monomorphic variable. It may also be helpful to explicitly indicate the underlying monomorphic symbol with a type subscript.

Note that the mixing of monomorphic and polymorphic syntax has been facilitated by using the same notation for variables in both the monomorphic and polymorphic cases. But the meanings in each case are somewhat different: in the monomorphic case, x stands for a variable of some specified type; in the polymorphic case, x takes the type inferred from its context. However, in the monomorphic case, since the syntax used for variables does not indicate their type, it effectively has to be inferred as well. Thus, while there is therefore potential for confusion about which kind of variable is intended in the mixed case, it does not really matter – in both cases, the type of the variable is inferred from its context (or, in extreme cases, specified using an explicit grounding type substitution).

B.1.11 Standard Predicates

In this section, grammars for defining spaces of predicates that appear in hypothesis languages in learning applications are studied. There are two parts to this: first a class of predicates, called standard predicates, that have a convenient form for use in grammars is defined; then a predicate grammar formalism, called a predicate rewrite system, that generates standard predicates is defined.

Predicates are built up by composing basic constants called transformations. Composition is handled by the (reverse) composition function

$$\diamond : (a \to b) \to (b \to c) \to (a \to c)$$

defined by $((f \diamond g)\ x) = (g\ (f\ x))$.

Definition B.1.50. A *transformation* f is a constant having a signature of the form

$$f : (\varrho_1 \to o) \to \cdots \to (\varrho_k \to o) \to \mu \to \sigma,$$

where any parameters in $\varrho_1, \ldots, \varrho_k$ and σ appear in μ, and $k \geq 0$. The type σ is called the *target* of the transformation. The number k is called the *rank* of the transformation.

Some examples of transformations are now given.

Example B.1.16. The transformation $top : a \to o$ is defined by $top\ x = \top$, for each x. The transformation $bottom : a \to o$ is defined by $bottom\ x = \bot$, for each x.

Example B.1.17. The transformation $\wedge_n : (a \to o) \to \cdots \to (a \to o) \to a \to o$ defined by

$$\wedge_n\ p_1 \ldots p_n\ x = (p_1\ x) \wedge \cdots \wedge (p_n\ x),$$

where $n \geq 2$, provides a 'conjunction' with n conjuncts.

The transformation $\vee_n : (a \to o) \to \cdots \to (a \to o) \to a \to o$ defined by

$$\vee_n\ p_1 \ldots p_n\ x = (p_1\ x) \vee \cdots \vee (p_n\ x),$$

where $n \geq 2$, provides a 'disjunction' with n disjuncts.

The transformation $\neg\ : (a \to o) \to a \to o$ defined by

$$\neg\ p\ x = \neg(p\ x),$$

provides negation.

Example B.1.18. The transformation $\mathit{setExists}_1 : (a \to o) \to \{a\} \to o$ defined by

$$\mathit{setExists}_1\ p\ t = \exists x.((p\ x) \wedge (x \in t)).$$

checks whether a set t has an element that satisfies p.

Example B.1.19. Given a type of the form $a_1 \times a_2 \times \cdots \times a_n$, one can define, for each i, a transformation $\mathit{proj}_i : a_1 \times a_2 \times \cdots \times a_n \to a_i$ defined by

$$\mathit{proj}_i\ (t_1, t_2, \ldots, t_n) = t_i$$

to project out the i-th component.

Example B.1.20. The constant $\mathit{isRectangle} : (\mathit{Float} \to o) \to (\mathit{Float} \to o) \to \mathit{Shape} \to o$ defined by

$$\mathit{isRectangle}\ p\ q\ t = \exists x.\exists y.((t = \mathit{Rectangle}\ x\ y) \wedge (p\ x) \wedge (q\ y))$$

is a transformation. For predicates p and q, $(\mathit{isRectangle}\ p\ q)$ is a predicate on geometrical shapes which returns $\top$ iff the shape is a rectangle whose length satisfies p and whose breadth satisfies q.

Next the definition of the class of predicates formed by composing transformations is presented. In the following definition, it is assumed that some (possibly infinite) class of transformations is given and all transformations considered are taken from this class. A standard predicate is defined by induction on the number of (occurrences of) transformations it contains as follows.

Notation. Let $\Box$ denote a (possibly empty) sequence of modalities $\Box_{j_1} \cdots \Box_{j_r}$.

Definition B.1.51. A *standard predicate* is a term of the form

$$\Box_1(f_1\ p_{1,1} \ldots p_{1,k_1}) \diamond \cdots \diamond \Box_n(f_n\ p_{n,1} \ldots p_{n,k_n}),$$

where f_i is a transformation of rank k_i $(i = 1, \ldots, n)$, the target of f_n is o, $\Box_i$ is a sequence of modalities $(i = 1, \ldots, n)$, p_{i,j_i} is a standard predicate $(i = 1, \ldots, n,\quad j_i = 1, \ldots, k_i)$, $k_i \geq 0$ $(i = 1, \ldots, n)$ and $n \geq 1$.

Example B.1.21. Let p and q be transformations of type $\sigma \to o$. Then

$$\boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \bullet\boldsymbol{B}_j p\ \blacklozenge\boldsymbol{B}_j q))$$

is a standard predicate of type $\{\sigma\} \to o$. If t is a (rigid) set of elements of type σ, then

$$(\boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \bullet\boldsymbol{B}_j p\ \blacklozenge\boldsymbol{B}_j q))\ \ t)$$

simplifies to

$$\boldsymbol{B}_i \exists x.((\bullet\boldsymbol{B}_j(p\ x) \wedge \blacklozenge\boldsymbol{B}_j(q\ x)) \wedge (x \in t)),$$

which is true iff agent i believes that there is an element x in t satisfying the property that at the last time agent j believed that x satisfied p and at some time in the past agent j believed that x satisfied q.

The set of all standard predicates is denoted by $\mathbf{S}$. Standard predicates have type of the form $\mu \to o$, for some type μ. Note that the set of standard predicates is defined *relative to* a previously declared set of transformations.

Definition B.1.52. For each $\alpha \in \mathfrak{S}^c$, define $\mathbf{S}_\alpha = \{p \in \mathbf{S} \mid p \text{ has type } \mu \to o \text{ and } \mu \text{ is more general than } \alpha\}$.

The intuitive meaning of $\mathbf{S}_\alpha$ is that it is the set of all predicates of a particular form given by the transformations on individuals of type α.

Careful attention needs to be paid to the subterms of a standard predicate, especially the subterms that are themselves standard predicates.

Definition B.1.53. Let $\Box_1(f_1\ \ p_{1,1} \ldots p_{1,k_1}) \diamond \cdots \diamond \Box_n(f_n\ \ p_{n,1} \ldots p_{n,k_n})$ be a standard predicate. A *suffix* of the standard predicate is a term of the form

$$\Box_i(f_i\ p_{i,1} \ldots p_{i,k_i}) \diamond \cdots \diamond \Box_n(f_n\ p_{n,1} \ldots p_{n,k_n}),$$

for some $i \in \{1, \ldots, n\}$. The suffix is *proper* if $i > 1$.

A *prefix* of the standard predicate is a term of the form

$$\Box_1(f_1\ p_{1,1} \ldots p_{1,k_1}) \diamond \cdots \diamond \Box_i(f_i\ p_{i,1} \ldots p_{i,k_i}),$$

for some $i \in \{1, \ldots, n\}$. The prefix is *proper* if $i < n$.

A suffix of a standard predicate is a standard predicate, but a prefix of a standard predicate may not be a predicate.

Proposition B.1.18. *Let p be a standard predicate*

$$\Box_1(f_1\ p_{1,1} \ldots p_{1,k_1}) \diamond \cdots \diamond \Box_n(f_n\ p_{n,1} \ldots p_{n,k_n})$$

and q a term. Then q is a subterm of p iff (at least) one of the following conditions holds.

1. *q is a suffix of p.*
2. *q is a subterm of $(\diamond\ \Box_i(f_i\ p_{i,1} \ldots p_{i,k_i}))$, for some $i \in \{1, \ldots, n-1\}$.*

3. q is a subterm of $\Box_n(f_n\ p_{n,1} \dots p_{n,k_n})$,

Proof. The proof is by induction on n. If $n = 1$, the result is obvious. Consider now a standard predicate

$$\Box_1(f_1\ p_{1,1} \dots p_{1,k_1}) \diamond \cdots \diamond \Box_{n+1}(f_{n+1}\ p_{n+1,1} \dots p_{n+1,k_{n+1}}),$$

which can be written in the form

$$((\diamond\ \Box_1(f_1\ p_{1,1} \dots p_{1,k_1}))\ \ \Box_2(f_2\ p_{2,1} \dots p_{2,k_2}) \diamond \cdots \diamond \Box_{n+1}(f_{n+1}\ p_{n+1,1} \dots p_{n+1,k_{n+1}})).$$

A subterm of this is either the term itself or a subterm of $(\diamond\ \Box_1(f_1\ p_{1,1} \dots p_{1,k_1}))$ or a subterm of $\Box_2(f_2\ p_{2,1} \dots p_{2,k_2}) \diamond \cdots \diamond \Box_{n+1}(f_{n+1}\ p_{n+1,1} \dots p_{n+1,k_{n+1}})$. The result follows by applying the induction hypothesis to this last term. □

Proposition B.1.19. *Let p be a standard predicate*

$$\Box_1(f_1\ p_{1,1} \dots p_{1,k_1}) \diamond \cdots \diamond \Box_n(f_n\ p_{n,1} \dots p_{n,k_n})$$

and q a subterm of p. Then q is a standard predicate iff (at least) one of the following conditions holds.

1. *q is a suffix of p.*
2. *q has the form $\Box_i(f_i\ p_{i,1} \dots p_{i,k_i})$ or $\Box_{i,2} \cdots \Box_{i,l_i}(f_i\ p_{i,1} \dots p_{i,k_i})$ or ... or $(f_i\ p_{i,1} \dots p_{i,k_i})$, for some $i \in \{1, \dots, n\}$, and f_i has target o, where $\Box_i$ is $\Box_{i,1} \cdots \Box_{i,l_i}$.*
3. *q is a subterm of $p_{i,j}$, for some i and j, and q is a standard predicate.*

Proof. According to Proposition B.1.18, the subterms of p are the suffixes of p, the subterms of $(\diamond\ \Box_i(f_i\ p_{i,1} \dots p_{i,k_i}))$, for $i = 1, \dots, n-1$, and the subterms of $\Box_n(f_n\ p_{n,1} \dots p_{n,k_n})$, Thus it is only necessary to establish which of these subterms are standard predicates. First, all the suffixes are standard predicates. Since $\diamond$ is not a standard predicate, it remains to investigate the subterms of $\Box_i(f_i\ p_{i,1} \dots p_{i,k_i})$, for $i = 1, \dots, n$. Now, provided the target of f_i is o, $\Box_i(f_i\ p_{i,1} \dots p_{i,k_i})$ and $\Box_{i,2} \cdots \Box_{i,l_i}(f_i\ p_{i,1} \dots p_{i,k_i})$ and ... and $(f_i\ p_{i,1} \dots p_{i,k_i})$ are standard predicates. Also the only proper subterms of $(f_i\ p_{i,1} \dots p_{i,k_i})$ that could possibly be standard predicates are subterms of $p_{i,1}, \dots, p_{i,k_i}$. (The reason is that for f_i to be the first symbol in a standard predicate, it must be followed by *all* k of its predicate arguments.) The result follows. □

Example B.1.22. Consider the standard predicate

$$(f_1\ p) \diamond (f_2\ q\ r) \diamond (f_3\ s),$$

where the transformations f_1, f_2, f_3, p, q, r, and s have suitable signatures. Thus p, q, r, and s are predicates. Suppose also that neither f_1 nor f_2 have target o. Taking the associativity into account, this standard predicate can be written more precisely as

$$((\diamond\ (f_1\ p))\ \ ((\diamond\ ((f_2\ q)\ r))\ \ (f_3\ s))).$$

The subterms of $(f_1\ p) \diamond (f_2\ q\ r) \diamond (f_3\ s)$ are illustrated in Figure B.2. The proper subterms that are standard predicates are $(f_2\ q\ r) \diamond (f_3\ s)$, $(f_3\ s)$, p, q, r, and s.

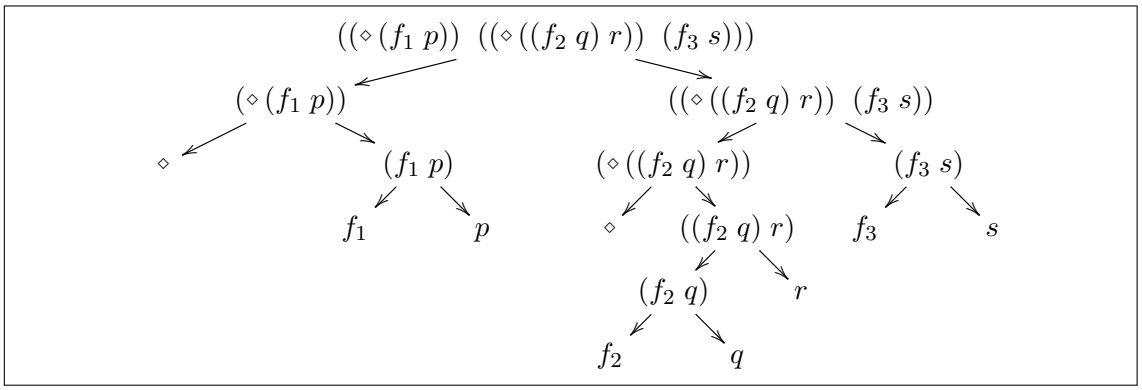

Figure B.2: Subterms of a standard predicate

The class of standard predicates just defined contains some redundancy in that there are syntactically distinct predicates that are equivalent, in a certain sense.

Definition B.1.54. The theory consisting of the definitions of the transformations (and definitions of associated constants) is called the *background theory*.

In an agent application, the background theory would normally be the belief base of an agent (or, perhaps, the collected belief bases of several agents).

Definition B.1.55. Let $\mathcal{B}$ be the background theory, and s and t terms having the same type (up to variants). Then s and t are *equivalent with respect to* $\mathcal{B}$ if $\Box(s = t)$ is a consequence of $\mathcal{B}$, for every sequence of modalities $\Box$.

Equivalence with respect to a background theory is an equivalence relation on the set of terms. When discussing equivalence in the following, explicit mention of the background theory is usually suppressed.

Example B.1.23. Without spelling out the background theory, one would expect that the standard predicates $(\wedge_2\ p\ q)$ and $(\wedge_2\ q\ p)$ are equivalent, where p and q are standard predicates. Similarly, $(\vee_2\ p\ q)$ and $(\vee_2\ q\ p)$ are equivalent.

It is important to remove as much of this redundancy in standard predicates as possible. Ideally, one would like to be able to determine just *one* representative from each class of equivalent predicates. However, determining the equivalent predicates is undecidable, so one usually settles for some easily checked syntactic conditions that reveal equivalence of predicates. Thus these syntactic conditions are sufficient, but not necessary, for equivalence. These considerations motivate the next definition.

Definition B.1.56. A transformation f is *symmetric* if it has a signature of the form

$$f : (\varrho \to o) \to \cdots \to (\varrho \to o) \to \mu \to \sigma,$$

and, whenever $(f\ p_1 \dots p_k)$ is a term, where $p_1, \dots, p_k$ are standard predicates, it follows that $(f\ p_1 \dots p_k)$ and $(f\ p_{i_1} \dots p_{i_k})$ are equivalent, for all permutations i of $\{1, \dots, k\}$, where k is the rank of f.

Note that if $(f\ p_1 \dots p_k)$ is a term, then $(f\ p_{i_1} \dots p_{i_k})$ is a term of the same type. It is common for a symmetric transformation to have o as its target; however, this does not have to be the case. Clearly every transformation of rank k, where $k \leq 1$, is (trivially) symmetric. Furthermore, the transformations $\wedge_n$ and $\vee_n$ are symmetric. However, *isRectangle* is not symmetric since it distinguishes the length argument from the breadth argument.

Since any permutation of the predicate arguments of a symmetric transformation produces an equivalent function, it is advisable to choose one particular order of arguments and ignore the others. For this purpose, a total order on standard predicates is defined and then arguments for symmetric transformations are chosen in increasing order according to this total order. To define the total order on standard predicates, one must start with a total order on modalities and transformations. Therefore, it is supposed that the modalities $\Box_1, \dots, \Box_m$ are ordered according to some (arbitrary) strict total order $<$. This order can can lifted to the strict total order also denoted by $<$ on (finite) sequences of modalities given by the lexicographic order. Furthermore, transformations are ordered according to some (arbitrary) strict total order also denoted by $<$.

In preparation for the definition of $\prec$, the total order on standard predicates, a structural result about standard predicates is needed.

Proposition B.1.20. *Let $p, q \in \mathbf{S}$, where p is*

$$\Box_{p,1}(f_1\ p_{1,1} \dots p_{1,k_1}) \diamond \cdots \diamond \Box_{p,n}(f_n\ p_{n,1} \dots p_{n,k_n})$$

and q is

$$\Box_{q,1}(g_1\ q_{1,1} \dots q_{1,s_1}) \diamond \cdots \diamond \Box_{q,r}(g_r\ q_{r,1} \dots q_{r,s_r}).$$

Suppose that p and q are (syntactically) distinct. Then exactly one of the following alternatives holds.

1. *One of p or q is a proper prefix of the other.*
2. *There exists i such that $\Box_{p,i}$ and $\Box_{q,i}$ are distinct, and p and q agree to the left of $\Box_{p,i}$ and $\Box_{q,i}$.*
3. *There exists i such that the transformation f_i in p and the transformation g_i in q are distinct, and p and q agree to the left of f_i and g_i.*
4. *There exist i and j such that $p_{i,j}$ in p and $q_{i,j}$ in q are distinct, and p and q agree to the left of $p_{i,j}$ and $q_{i,j}$.*

Proof. The proof is by induction on the maximum of the number of (occurrences of) transformations in p and the number in q.

If this maximum number is one, then the result is obvious with the second and third alternatives being the only possibilities.

Now consider the inductive step. Suppose first that $\Box_{p,1}$ and $\Box_{q,1}$ are distinct. Then the second alternative holds. Suppose next that $\Box_{p,1}$ and $\Box_{q,1}$ are identical, but f_1 is distinct from g_1. Then the third alternative holds. Otherwise, f_1 and g_1 are identical, which gives rise to two cases: either the arguments to f_1 in p and g_1 in q are pairwise

identical or they are not. In the first case, consider the suffixes of p and q obtained by removing the common prefix $\Box_{p,1}(f_1\ p_{1,1} \ldots p_{1,k_1})$. If one of the suffixes is empty, then the first alternative holds; otherwise, the inductive hypothesis can be applied to the suffixes to obtain the result. In the other case, the leftmost pair of $p_{1,j}$ and $q_{1,j}$ that are distinct gives the fourth alternative. □

Example B.1.24. As an illustration of the first alternative in the preceding proposition, consider $\Box_1 p$ and $\Box_1 p \diamond \neg$, for some standard predicate p. For the second alternative, consider $\Box_1 p$ and $\Box_2 p$, for some standard predicate p. For the fourth alternative, consider $\Box_1(\wedge_2\ p\ q)$ and $\Box_1(\wedge_2\ (p \diamond \neg)\ q)$, for some standard predicates p and q.

The following definition of the relation $p \prec q$ uses induction on the maximum of the number of (occurrences of) transformations in p and the number in q. To emphasize: the definition of $\prec$ depends upon the order on the modalities and the order on the transformations.

Definition B.1.57. The binary relation $\prec$ on $\mathbf{S}$ is defined inductively as follows. Let $p, q \in \mathbf{S}$, where p is

$$\Box_{p,1}(f_1\ p_{1,1} \ldots p_{1,k_1}) \diamond \cdots \diamond \Box_{p,n}(f_n\ p_{n,1} \ldots p_{n,k_n})$$

and q is

$$\Box_{q,1}(g_1\ q_{1,1} \ldots q_{1,s_1}) \diamond \cdots \diamond \Box_{q,r}(g_r\ q_{r,1} \ldots q_{r,s_r}).$$

Then $p \prec q$ if one of the following holds.

1. p is a proper prefix of q.
2. There exists an i such that $\Box_{p,i} < \Box_{q,i}$ and p and q agree to the left of $\Box_{p,i}$ and $\Box_{q,i}$.
3. There exists i such that $f_i < g_i$, and p and q agree to the left of f_i and g_i.
4. There exist i and j such that $p_{i,j} \prec q_{i,j}$, and p and q agree to the left of $p_{i,j}$ and $q_{i,j}$.

Proposition B.1.21. *The relation $\prec$ is a strict total order on $\mathbf{S}$.*

Proof. Let $p, q, r \in \mathbf{S}$, where p is

$$\Box_{p,1}(f_1\ p_{1,1} \ldots p_{1,k_1}) \diamond \cdots \diamond \Box_{p,n}(f_n\ p_{n,1} \ldots p_{n,k_n}),$$

q is

$$\Box_{q,1}(g_1\ q_{1,1} \ldots q_{1,s_1}) \diamond \cdots \diamond \Box_{q,u}(g_u\ q_{u,1} \ldots q_{u,s_u})$$

and r is

$$\Box_{r,1}(h_1\ r_{1,1} \ldots r_{1,t_1}) \diamond \cdots \diamond \Box_{r,v}(h_v\ r_{v,1} \ldots r_{v,t_v}).$$

First, it is shown by induction on the number of (occurrences of) transformations in p that $p \not\prec p$. For this, note that p cannot be a proper prefix of itself. Also there cannot exist i such that $\Box_{p,i} < \Box_{p,i}$, since the lexicographic order is a strict total order on sequences

of modalities, and there cannot exist i such that $f_i < f_i$, since $<$ is a strict total order. Furthermore, there cannot exist i and j such that $p_{i,j} \prec p_{i,j}$, by the induction hypothesis.

Next it is shown by induction on the maximum of the number of transformations in p and the number in q that $p \prec q$ implies $q \not\prec p$. Thus suppose that $p \prec q$. If p is a proper prefix of q, then it is clear that $q \not\prec p$. If there exists an i such that $\Box_{p,i} < \Box_{q,i}$ and p and q agree to the left of $\Box_{p,i}$ and $\Box_{q,i}$, then it is clear that $q \not\prec p$, since the lexicographic order $<$ is a strict total order. If there exists i such that $f_i < g_i$, and p and q agree to the left of f_i and g_i, then it is clear that $q \not\prec p$, since $<$ is a strict total order on transformations. Finally, if there exist i and j such that $p_{i,j} \prec q_{i,j}$, and p and q agree to the left of $p_{i,j}$ and $q_{i,j}$, then $q \not\prec p$ since, by the induction hypothesis, it follows that $q_{i,j} \not\prec p_{i,j}$.

Now it is shown by induction on the maximum of the number of transformations in p, the number in q, and the number in r that $p \prec q$ and $q \prec r$ imply that $p \prec r$.

For the base step, each of p, q and r contains a single transformation. If the sequences of modalities for the three standard predicates are not all the same, then the result follows immediately from the transitivity of the lexicographic order on sequences of modalities. Otherwise, the result follows immediately from the transitivity of the order on transformations.

Now the inductive step is considered. There are four cases to consider, corresponding to the four cases in the definition of $p \prec q$.

Suppose first that p is a proper prefix of q. If q is a proper prefix of r, then p is a proper prefix of r and so $p \prec r$. If there exists i such that such that $\Box_{q,i} < \Box_{r,i}$ and q and r agree to the left of $\Box_{q,i}$ and $\Box_{r,i}$, then either p is a proper prefix of r or $\Box_{p,i} < \Box_{r,i}$ and p and r agree to the left of $\Box_{p,i}$ and $\Box_{r,i}$, so that $p \prec r$. If there exists i such that $g_i < h_i$, and q and r agree to the left of g_i and h_i, then either p is a proper prefix of r or $f_i < h_i$ and p and r agree to the left of f_i and h_i, and so $p \prec r$. If there exist i and j such that $q_{i,j} \prec r_{i,j}$, and q and r agree to the left of $q_{i,j}$ and $r_{i,j}$, then either p is a proper prefix of r or $p_{i,j} \prec r_{i,j}$ and p and r agree to the left of $p_{i,j}$ and $r_{i,j}$, and so $p \prec r$.

For the second case, suppose there exists an i such that $\Box_{p,i} < \Box_{q,i}$ and p and q agree to the left of $\Box_{p,i}$ and $\Box_{q,i}$. If q is a proper prefix of r, then $\Box_{p,i} < \Box_{r,i}$ and p and r agree up to this point, so that $p \prec r$. If there exists an i' such that $\Box_{q,i'} < \Box_{r,i'}$ and q and r agree to the left of $\Box_{q,i'}$ and $\Box_{r,i'}$, then if $i \neq i'$, clearly it is the case that $p \prec r$, and if $i = i'$, then $p \prec r$, by the transitivity of the lexicographic order $<$ on sequences of modalities. If there exists i' such that $g_{i'} < h_{i'}$, and q and r agree to the left of $g_{i'}$ and $h_{i'}$, then if $i > i'$, then $f_{i'} < h_{i'}$ and p and r agree up to this point, so that $p \prec r$, and if $i \leq i'$, then $\Box_{p,i} < \Box_{r,i}$ and p and r agree up to this point, so that $p \prec r$. If there exist i' and j such that $q_{i',j} \prec r_{i',j}$, and q and r agree to the left of $q_{i',j}$ and $r_{i',j}$, then if $i > i'$, then $p_{i',j} \prec r_{i',j}$, and p and r agree up to this point, so that $p \prec r$, and if $i \leq i'$, then $\Box_{p,i} < \Box_{r,i}$ and p and r agree up to this point, so that $p \prec r$.

For the third case, suppose there exists i such that $f_i < g_i$, and p and q agree to the left of f_i and g_i. If q is a proper prefix of r, then $f_i < h_i$, and p and r agree to the left of f_i and h_i, and so $p \prec r$. If there exists i' such that $\Box_{q,i'} < \Box_{r,i'}$ and q and r agree to the left of $\Box_{q,i'}$ and $\Box_{r,i'}$, then, if $i \geq i'$, then $\Box_{p,i'} < \Box_{r,i'}$ and p and r agree to the left of $\Box_{p,i'}$ and $\Box_{r,i'}$, so that $p \prec r$, and if $i < i'$, then $f_i < h_i$, and p and r agree to the left of f_i and h_i, so that $p \prec r$. If there exists i' such that $g_{i'} < h_{i'}$, and q and r agree to the left of $g_{i'}$ and $h_{i'}$, then, for $i'' = min(i, i')$, $f_{i''} < h_{i''}$ and p and r agree to the left of $f_{i''}$ and $h_{i''}$, and so $p \prec r$. Finally, suppose that there exist i' and j such that $q_{i',j} \prec r_{i',j}$, and q and r

agree to the left of $q_{i',j}$ and $r_{i',j}$. If $i \leq i'$, then $f_i < h_i$ and p and r agree to the left of f_i and h_i. If $i' < i$, then $p_{i',j} \prec r_{i',j}$ and p and r agree to the left of $p_{i',j}$ and $r_{i',j}$. In either case, $p \prec r$.

For the fourth case, suppose there exist i and j such that $p_{i,j} \prec q_{i,j}$, and p and q agree to the left of $p_{i,j}$ and $q_{i,j}$. If q is a proper prefix of r, then $p_{i,j} \prec r_{i,j}$, and p and r agree to the left of $p_{i,j}$ and $r_{i,j}$, and so $p \prec r$. If there exists i' such that $\Box_{q,i'} < \Box_{r,i'}$ and q and r agree to the left of $\Box_{q,i'}$ and $\Box_{r,i'}$, then if $i \geq i'$, then $\Box_{p,i'} < \Box_{r,i'}$ and p and r agree to the left of $\Box_{p,i'}$ and $\Box_{r,i'}$, so that $p \prec r$, and if $i < i'$, then $p_{i,j} \prec r_{i,j}$, and p and r agree to the left of $p_{i,j}$ and $r_{i,j}$, so that $p \prec r$. Suppose that there exists i' such that $g_{i'} < h_{i'}$, and q and r agree to the left of $g_{i'}$ and $h_{i'}$. If $i' \leq i$, then $f_{i'} < h_{i'}$ and p and r agree to the left of $f_{i'}$ and $h_{i'}$. If $i < i'$, then $p_{i,j_i} \prec r_{i,j_i}$ and p and r agree to the left of p_{i,j_i} and r_{i,j_i}. In either case, $p \prec r$. Finally, suppose there exist i' and j' such that $q_{i',j'} \prec r_{i',j'}$, and q and r agree to the left of $q_{i',j'}$ and $r_{i',j'}$. Put $i'' = min(i, i')$ and let j'' be j if $i < i'$, or j' if $i' < i$, or $min(j, j')$, otherwise. Then $p_{i'',j''} \prec r_{i'',j''}$ and p and r agree to the left of $p_{i'',j''}$ and $r_{i'',j''}$, and so $p \prec r$. (Here the induction hypothesis is used if $i = i'$ and $j = j'$.)

Thus $\prec$ is a strict partial order.

Finally, it is shown that $\prec$ is total, that is, for any standard predicates p and q, either $p = q$ or $p \prec q$ or $q \prec p$. The proof proceeds by induction on the maximum of the number of transformations in p and the number in q. The base step is straightforward. Now the inductive step is considered. Suppose that p and q are distinct. By Proposition B.1.20, either one of p or q is a proper prefix of the other; or there exists i such that $\Box_{p,i}$ and $\Box_{q,i}$ are distinct, and p and q agree to the left of $\Box_{p,i}$ and $\Box_{q,i}$; or there exists i such that the transformation f_i in p and the transformation g_i in q are distinct, and p and q agree to the left of f_i and g_i; or there exist i and j such that $p_{i,j}$ in p and $q_{i,j}$ in q are distinct, and p and q agree to the left of $p_{i,j}$ and $q_{i,j}$. In the first case, either $p \prec q$ or $q \prec p$. In the second case, either $p \prec q$ or $q \prec p$, since the lexicographic ordering on sequences of modalities is a total order. In the third case, since $<$ is a total order on transformations, either $f_i < g_i$ or $g_i < f_i$, and hence either $p \prec q$ or $q \prec p$. In the fourth case, by the induction hypothesis, either $p_{i,j} \prec q_{i,j}$ or $q_{i,j} \prec p_{i,j}$, and so once again either $p \prec q$ or $q \prec p$. □

The relation $\preceq$ is defined by $p \preceq q$ if either $p = q$ or $p \prec q$. Clearly, $\preceq$ is a total order on $\mathbf{S}$.

Now the class of regular predicates can be defined by induction on the number of transformations in a predicate.

Definition B.1.58. A standard predicate

$$\Box_1(f_1\ p_{1,1} \dots p_{1,k_1}) \diamond \cdots \diamond \Box_n(f_n\ p_{n,1} \dots p_{n,k_n})$$

is *regular* if p_{i,j_i} is a regular predicate, for $i = 1, \dots, n$ and $j_i = 1, \dots, k_i$, and f_i is symmetric implies that $p_{i,1} \preceq \cdots \preceq p_{i,k_i}$, for $i = 1, \dots, n$.

The set of all regular predicates is denoted by $\mathbf{R}$.

The next result shows that each standard predicate is equivalent to a regular predicate and hence attention can be confined to the generally much smaller class of regular predicates.

Proposition B.1.22. *For every $p \in \mathbf{S}$, there exists $q \in \mathbf{R}$ such that p and q are equivalent.*

Proof. The existence of the regular predicate q is shown by induction on the number of transformations in p. Let p be

$$\Box_1(f_1\ p_{1,1} \dots p_{1,k_1}) \diamond \cdots \diamond \Box_n(f_n\ p_{n,1} \dots p_{n,k_n}).$$

By the induction hypothesis, for each $i = 1, \dots, n$ and $j_i = 1, \dots, k_i$, there is a regular predicate p'_{i,j_i} such that p'_{i,j_i} and p_{i,j_i} are equivalent. Let r be $\Box_1(f_1\ p'_{1,1} \dots p'_{1,k_1}) \diamond \cdots \diamond \Box_n(f_n\ p'_{n,1} \dots p'_{n,k_n})$. Then p and r are equivalent. Now, since $\preceq$ is a total order on standard predicates, for each symmetric transformation f_i ($i \in \{1, \dots, n\}$) in r, one can sort the arguments $p'_{i,1}, \dots, p'_{i,k_i}$ according to $\preceq$ to obtain the desired regular predicate q such that p and q are equivalent. □

B.1.12 Predicate Rewrite Systems

This section gives the definition of the concept of a predicate rewrite system.

First, this is explained informally. A predicate rewrite is an expression of the form $p \rightarrowtail q$, where p and q are standard predicates. The predicate p is called the *head* and q is the *body* of the rewrite. A predicate rewrite system is a finite set of predicate rewrites. One should think of a predicate rewrite system as a kind of grammar for generating a particular class of predicates. Roughly speaking, this works as follows. Starting from the weakest predicate *top*, all predicate rewrites that have *top* (of the appropriate type) in the head are selected to make up child predicates that consist of the bodies of these predicate rewrites. Then, for each child predicate and each redex in that predicate, all child predicates are generated by replacing each redex by the body of the predicate rewrite whose head is identical to the redex. This generation of predicates continues to produce the entire space of predicates given by the predicate rewrite system.

Here is the formal development.

Definition B.1.59. A *predicate rewrite system* is a finite relation $\rightarrowtail$ on $\mathbf{S}$ satisfying the following two properties.

1. For each $p \rightarrowtail q$, the type of p is more general than the type of q.
2. For each $p \rightarrowtail q$, there does not exist $s \rightarrowtail t$ such that q is a proper subterm of s.

If $p \rightarrowtail q$, then $p \rightarrowtail q$ is called a *predicate rewrite*, p the *head*, and q the *body* of the predicate rewrite.

The second condition of Definition B.1.59 states that no body of a rewrite is a proper subterm of the head of any rewrite. In practice, the heads of rewrites are usually standard predicates consisting of just one transformation, such as *top*. In this case, the second condition is automatically satisfied, since the heads have no proper subterms at all.

Definition B.1.60. A subterm of a standard predicate

$$\Box_1(f_1\ p_{1,1} \dots p_{1,k_1}) \diamond \cdots \diamond \Box_n(f_n\ p_{n,1} \dots p_{n,k_n})$$

is *eligible* if it is a suffix of the standard predicate or it is an eligible subterm of $p_{i,j}$, for some $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, k_i\}$.

Example B.1.25. The eligible subterms of

$$\Box_1 \mathit{vertices} \diamond \Box_1 \Box_2 (\mathit{setExists}_2\ (\wedge_2\ (\mathit{proj}_1 \diamond (= A))\ (\mathit{proj}_2 \diamond (= 3)))\ (\mathit{proj}_1 \diamond (= B)))$$

are

$$\Box_1 \mathit{vertices} \diamond \Box_1 \Box_2 (\mathit{setExists}_2\ (\wedge_2\ (\mathit{proj}_1 \diamond (= A))\ (\mathit{proj}_2 \diamond (= 3)))\ (\mathit{proj}_1 \diamond (= B))),$$
$$\Box_1 \Box_2 \mathit{setExists}_2\ (\wedge_2\ (\mathit{proj}_1 \diamond (= A))\ (\mathit{proj}_2 \diamond (= 3)))\ (\mathit{proj}_1 \diamond (= B)),$$
$$\wedge_2\ (\mathit{proj}_1 \diamond (= A))\ (\mathit{proj}_2 \diamond (= 3)),$$
$$\mathit{proj}_1 \diamond (= A),$$
$$\mathit{proj}_2 \diamond (= 3),$$
$$(= A),$$
$$(= 3),$$
$$\mathit{proj}_1 \diamond (= B), \text{ and}$$
$$(= B).$$

Proposition B.1.23. *An eligible subterm of a standard predicate is a standard predicate.*

Proof. This is a straightforward induction argument. □

Proposition B.1.24. *Let p, q, and r be standard predicates, and q an eligible subterm of p such that $p[q/r]$ is a term. Then $p[q/r]$ is a standard predicate.*

Proof. The proof is by induction on the number of transformations in p. If the number of transformations in p is one, then the result is obvious. Suppose the result holds for standard predicates that have $< m$ transformations and p has m transformations. Let p have the form

$$\Box_1 (f_1\ p_{1,1} \dots p_{1,k_1}) \diamond \cdots \diamond \Box_n (f_n\ p_{n,1} \dots p_{n,k_n}).$$

If q is a suffix of p, then the result follows since $p[q/r]$ is a term. If q is an eligible subterm of some p_{i,j_i}, then the induction hypothesis gives that $p_{i,j_i}[q/r]$ is a standard predicate. Since $p[q/r]$ is a term, it follows that $p[q/r]$ is a standard predicate. □

Definition B.1.61. Let $\rightarrowtail$ be a predicate rewrite system and p a standard predicate. An eligible subterm r of p is a *redex* with respect to $\rightarrowtail$ if there exists a predicate rewrite $r \rightarrowtail b$ such that $p[r/b]$ is a standard predicate. In this case, r is said to be a redex *via* $r \rightarrowtail b$.

It follows from Proposition B.1.24 that r is a redex iff $p[r/b]$ is a term. The phrase 'redex with respect to $\rightarrowtail$' is usually abbreviated to simply 'redex' with the predicate rewrite system understood.

Definition B.1.62. Let $\rightarrowtail$ be a predicate rewrite system, and p and q standard predicates. Then q is obtained by a *predicate derivation step* from p using $\rightarrowtail$ if there is a redex r via $r \rightarrowtail b$ in p and $q = p[r/b]$. The redex r is called the *selected* redex.

Definition B.1.63. A *predicate derivation* with respect to a predicate rewrite system $\rightarrowtail$ is a finite sequence $\langle p_0, p_1, \ldots, p_n \rangle$ of standard predicates such that p_i is obtained by a derivation step from p_{i-1} using $\rightarrowtail$, for $i = 1, \ldots, n$. The *length* of the predicate derivation is n. The standard predicate p_0 is called the *initial predicate* and the standard predicate p_n is called the *final predicate*.

Usually the initial predicate is *top*, the weakest predicate.

Example B.1.26. Consider the following predicate rewrite system.

$$
\begin{aligned}
&\mathit{top} \rightarrowtail \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \mathit{top}\ \mathit{top}))\\
&\mathit{top} \rightarrowtail \bullet\boldsymbol{B}_j\mathit{top}\\
&\mathit{top} \rightarrowtail \blacklozenge\boldsymbol{B}_j\mathit{top}\\
&\mathit{top} \rightarrowtail p\\
&\mathit{top} \rightarrowtail q\\
&\mathit{top} \rightarrowtail r.
\end{aligned}
$$

The following is a derivation in the predicate space defined by the rewrite system.

$$
\begin{aligned}
&\mathit{top}\\
&\rightsquigarrow\ \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \mathit{top}\ \mathit{top}))\\
&\rightsquigarrow\ \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \bullet\boldsymbol{B}_j\mathit{top}\ \ \mathit{top}))\\
&\rightsquigarrow\ \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \bullet\boldsymbol{B}_j p\ \ \mathit{top}))\\
&\qquad\vdots\\
&\rightsquigarrow\ \boldsymbol{B}_i(\mathit{setExists}_1\ (\wedge_2\ \bullet\boldsymbol{B}_j p\ \blacklozenge\boldsymbol{B}_j q)).
\end{aligned}
$$

The set $P_{\rightarrowtail}$ of predicates that can be generated from a predicate rewrite system $\rightarrowtail$ is called a *predicate language*. Given some predicate language, it remains to specify the *hypothesis language*, that is, the form of learned theories that employ predicates in the predicate language.

A language of particular importance for belief acquisition is defined next.

Definition B.1.64. Let α be a type. A *basic predicate* for the type α is one of the form $(= t)$, for some $t \in \mathfrak{B}_\alpha$.

The set $\mathbf{B}_\alpha = \{(= t) \mid t \in \mathfrak{B}_\alpha\}$ of basic predicates for the type α is called the *basic language* for the type α.

Note that $\mathbf{B}_\alpha$ is the set of predicates that can be generated from the predicate rewrite system consisting of all rewrites of the form

$$\mathit{top} \rightarrowtail (= t),$$

where $t \in \mathfrak{B}_\alpha$.

B.2 Semantics

This section gives the model theory for the logic. The semantic notions of interpretation, denotation, validity, and consequence are presented.

B.2.1 Interpretations

Definition B.2.1. A *domain set* for an alphabet is a collection $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$ of non-empty sets satisfying the following conditions.

1. $\mathcal{D}_o = \{\mathsf{T}, \mathsf{F}\}$.
2. For all $\alpha, \beta \in \mathfrak{S}$, $\mathcal{D}_{\alpha\to\beta}$ is some collection of functions from $\mathcal{D}_\alpha$ to $\mathcal{D}_\beta$.
3. For all $\alpha_1, \ldots, \alpha_n \in \mathfrak{S}$, $\mathcal{D}_{\alpha_1\times\cdots\times\alpha_n}$ is the cartesian product $\mathcal{D}_{\alpha_1} \times \cdots \times \mathcal{D}_{\alpha_n}$.

Each $\mathcal{D}_\alpha$ is called a *domain*.

Definition B.2.2. A *frame* is a pair $\langle W, \{R_i\}_{i=1}^m\rangle$ consisting of a non-empty set W and a set of binary relations $\{R_i\}_{i=1}^m$ on W.

Each element of W is called a *world* and each relation R_i is called an *accessibility* relation.

Definition B.2.3. An *augmented frame* for an alphabet is a triple $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}\rangle$, where $\langle W, \{R_i\}_{i=1}^m\rangle$ is a frame and $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$ is a domain set for the alphabet.

Definition B.2.4. A *valuation* with respect to an augmented frame $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}\rangle$ is a mapping V that maps each pair consisting of a constant having signature α and a world in W to an element of $\mathcal{D}_\alpha$ (called the *denotation* of the constant in the world) such that if a constant C is rigid, then $V(C, w) = V(C, w')$, for each $w, w' \in W$.

Definition B.2.5. An *interpretation* for an alphabet is a quadruple $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$, where $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}\rangle$ is an augmented frame for the alphabet and V is a valuation with respect to this augmented frame such that the following conditions are satisfied.

1. For all $w \in W$, $V(\top, w) = \mathsf{T}$ and $V(\bot, w) = \mathsf{F}$.
2. For $= : \alpha \to \alpha \to o$ and all $w \in W$, $V(=, w)$ is the element of $\mathcal{D}_{\alpha\to\alpha\to o}$ defined by

$$V(=, w)\ x\ y = \begin{cases} \mathsf{T} & \text{if } x = y \\ \mathsf{F} & \text{otherwise,} \end{cases}$$

   for all $x, y \in \mathcal{D}_\alpha$.
3. For $\neg : \alpha \to \alpha$ and all $w \in W$, $V(\neg, w)$ is the element of $\mathcal{D}_{\alpha\to\alpha}$ defined by

$$V(\neg, w)\ f\ d_1\ \ldots\ d_n = \begin{cases} \mathsf{T} & \text{if } f\ d_1\ \ldots\ d_n = \mathsf{F} \\ \mathsf{F} & \text{otherwise,} \end{cases}$$

   for all $f \in \mathcal{D}_\alpha$ and $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, where $\alpha \triangleq \alpha_1 \to \cdots \to \alpha_n \to o$, and $n \geq 0$.

4. For $\wedge : \alpha \to \alpha \to \alpha$ and all $w \in W$, $V(\wedge, w)$ is the element of $\mathcal{D}_{\alpha\to\alpha\to\alpha}$ defined by

$$V(\wedge, w)\ f\ g\ d_1\ \ldots\ d_n = \begin{cases} \mathsf{T} & \text{if } f\ d_1\ \ldots\ d_n = \mathsf{T} \text{ and } g\ d_1\ \ldots\ d_n = \mathsf{T} \\ \mathsf{F} & \text{otherwise,} \end{cases}$$

for all $f, g \in \mathcal{D}_\alpha$ and $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, where $\alpha \triangleq \alpha_1 \to \cdots \to \alpha_n \to o$, and $n \geq 0$.

5. For $\vee : \alpha \to \alpha \to \alpha$ and all $w \in W$, $V(\vee, w)$ is the element of $\mathcal{D}_{\alpha\to\alpha\to\alpha}$ defined by

$$V(\vee, w)\ f\ g\ d_1\ \ldots\ d_n = \begin{cases} \mathsf{T} & \text{if } f\ d_1\ \ldots\ d_n = \mathsf{T} \text{ or } g\ d_1\ \ldots\ d_n = \mathsf{T} \\ \mathsf{F} & \text{otherwise,} \end{cases}$$

for all $f, g \in \mathcal{D}_\alpha$ and $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, where $\alpha \triangleq \alpha_1 \to \cdots \to \alpha_n \to o$, and $n \geq 0$.

6. For $\longrightarrow : \alpha \to \alpha \to \alpha$ and all $w \in W$, $V(\longrightarrow, w)$ is the element of $\mathcal{D}_{\alpha\to\alpha\to\alpha}$ defined by

$$V(\longrightarrow, w)\ f\ g\ d_1\ \ldots\ d_n = \begin{cases} \mathsf{T} & \text{if } f\ d_1\ \ldots\ d_n = \mathsf{F} \text{ or } g\ d_1\ \ldots\ d_n = \mathsf{T} \\ \mathsf{F} & \text{otherwise,} \end{cases}$$

for all $f, g \in \mathcal{D}_\alpha$ and $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, where $\alpha \triangleq \alpha_1 \to \cdots \to \alpha_n \to o$, and $n \geq 0$.

7. For $\Sigma : \alpha \to o$ and all $w \in W$, $V(\Sigma, w)$ is the element of $\mathcal{D}_{\alpha\to o}$ defined by

$$V(\Sigma, w)\ f = \begin{cases} \mathsf{T} & \text{if } f \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \\ \mathsf{F} & \text{otherwise,} \end{cases}$$

for all $f \in \mathcal{D}_\alpha$, where α is a biterm type of rank n.

8. For $\Pi : \alpha \to o$ and all $w \in W$, $V(\Pi, w)$ is the element of $\mathcal{D}_{\alpha\to o}$ defined by

$$V(\Pi, w)\ f = \begin{cases} \mathsf{T} & \text{if } f = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \\ \mathsf{F} & \text{otherwise,} \end{cases}$$

for all $f \in \mathcal{D}_\alpha$, where α is a biterm type of rank n.

In effect, the booleans, connectives, and quantifiers are given their usual fixed meanings in each world. For other constants, their meaning may change from world to world.

Note that the constant domain semantics is being used here, in contrast to the varying domain semantics. It may seem restrictive to employ the constant domain semantics but it turns out that, for many applications, the issues of keeping track of function definitions that vary over time and vary from agent to agent completely overshadow the subtleties of whether objects may come into, or go out of, existence. In other words, for many applications, the simpler constant domain semantics easily suffices. In any case, it can be shown that each semantics can simulate the other.

Definition B.2.6. A *pointed interpretation* is a pair of the form (I, w), where $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ is an interpretation and w a world in W.

B.2.2 Denotations

The next task is to define the denotation of a term with respect to an interpretation, world, and variable assignment. This is rather standard except for the case of modal terms. The discussion begins with some intuition that shows how to proceed in this case.

If t is a formula, then the meaning of $\Box_i t$ in a world is T if the meaning of t in all accessible worlds is T, its meaning is F if the meaning of t in all accessible worlds is F, and, in the other cases, the meaning of $\Box_i t$ is conventionally defined to be F. This suggests an obvious extension to terms t whose type has rank 0: if t has the same meaning in all accessible worlds, then the meaning of $\Box_i t$ should be this common meaning; otherwise, the meaning of $\Box_i t$ should be some default value. This definition then becomes the base case of an inductive definition on the rank of the type of t of the semantics of a modal term $\Box_i t$. Here is the machinery to do this.

Definition B.2.7. Let $\{X_i\}_{i=0}^n$ $(n \geq 0)$ be a family of non-empty sets, and d and e two distinguished elements in X_0 that are called the *default values*. Let F be a set of functions, where

$$f : X_n \to (X_{n-1} \to (X_{n-2} \to \cdots \to X_0) \cdots),$$

for each $f \in F$. (If $n = 0$, the meaning is that each f belongs to X_0.) Define

$$\mathcal{M}(F) : X_n \to (X_{n-1} \to (X_{n-2} \to \cdots \to X_0) \cdots)$$

as follows:

1. If F is empty, then $\mathcal{M}(F) = \lambda x_1. \ldots \lambda x_n.d$.
2. (a) If F is non-empty and $n = 0$, then $\mathcal{M}(F) \in X_0$ is defined by

$$\mathcal{M}(F) = \begin{cases} x & \text{if there exists } x \in X_0, \text{ for each } f \in F,\ f = x \\ e & \text{otherwise.} \end{cases}$$

   (b) If F is non-empty and $n > 0$, then $\mathcal{M}(F) : X_n \to (X_{n-1} \to (X_{n-2} \to \cdots \to X_0) \cdots)$ is defined by

$$\mathcal{M}(F)(x) = \mathcal{M}(\{f(x)\}_{f \in F}),$$

   for each $x \in X_n$.

If there exists $f \in F$ such that $f = g$, for all $g \in F$, then clearly $\mathcal{M}(F) = f$. Also, if $F = \bigcup_{i \in I} F_i$, then $\mathcal{M}(F) = \mathcal{M}(\{\mathcal{M}(F_i)\}_{i \in I})$.

The function $\mathcal{M}$ will be used to give meaning to modal terms.

Definition B.2.8. Let α be a type of rank 0. Then the *default values* in $\mathcal{D}_\alpha$ are two distinguished elements d and e in $\mathcal{D}_\alpha$. In the case of $\mathcal{D}_o$, d is T and e is F.

There are no restrictions on the choice of default values, other than the default values for $\mathcal{D}_o$ of T and F. This choice is made for consistency with the standard semantics of modalities applied to formulas.

Definition B.2.9. A *variable assignment* with respect to a domain set $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$ is a mapping that maps each variable of type α to an element of $\mathcal{D}_\alpha$.

Now everything is in place to give the denotation of a term with respect to an interpretation, world and variable assignment.

Definition B.2.10. Let t be a term, $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, and w a world in W. Then the *denotation* $\mathcal{V}(t, I, w, \nu)$ of t with respect to I, w and ν is defined inductively as follows.

1. $\mathcal{V}(x, I, w, \nu) = \nu(x)$, where x is a variable.
2. $\mathcal{V}(C, I, w, \nu) = V(C, w)$, where C is a constant.
3. $\mathcal{V}(\lambda x.s, I, w, \nu) =$ the function whose value for each $d \in \mathcal{D}_\alpha$ is $\mathcal{V}(s, I, w, \nu')$, where x has type α and ν' is ν except $\nu'(x) = d$.
4. $\mathcal{V}((s\ r), I, w, \nu) = \mathcal{V}(s, I, w, \nu)(\mathcal{V}(r, I, w, \nu))$.
5. $\mathcal{V}((t_1, \ldots, t_n), I, w, \nu) \ = \ (\mathcal{V}(t_1, I, w, \nu), \ldots, \mathcal{V}(t_n, I, w, \nu))$.
6. $\mathcal{V}(\Box_i t, I, w, \nu) = \mathcal{M}(\{\mathcal{V}(t, I, w', \nu)\}_{w'\in W'})$, where $W' = \{w' : w\, R_i\, w'\}$.

It is understood in Definition B.2.10 that there are enough functions in each $\mathcal{D}_{\alpha\to\beta}$ so that every term of type $\alpha \to \beta$ does have a denotation.

If the accessibility relation R_i is serial (that is, for all $w \in W$, there exists $w' \in W$ such that $w\, R_i\, w'$), then $\{\mathcal{V}(t, I, w', \nu)\}_{w'\in W'}$ in Part 6 of Definition B.2.10 is not empty.

A special case of Part 6 of Definition B.2.10 is when the type of t is o (that is, t is a formula). Then the semantics for $\Box_i t$ given by Condition 6 is exactly the same as the standard semantics for a modal formula.

In the case where there is a unique $w' \in W$ such that $w\, R_i\, w'$, it follows that $\mathcal{V}(\Box_i t, I, w, \nu) = \mathcal{V}(t, I, w', \nu)$.

Definition B.2.10 is an inductive construction employing Proposition B.4.4. To set this up, let the interpretation $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be fixed throughout. Then Definition B.2.10 defines a function

$$T : \mathfrak{L} \times W \times \mathcal{A} \to \mathfrak{D}$$

such that $T(t, w, \nu) = \mathcal{V}(t, I, w, \nu)$, for all $t \in \mathfrak{L}$, $w \in W$, and $\nu \in \mathcal{A}$, where $\mathcal{A}$ is the set of all variable assignments and $\mathfrak{D}$ is the (disjoint) union of the $\mathfrak{D}_\alpha$, for all $\alpha \in \mathfrak{S}$. Consider now the substring relation $\prec$ on $\mathfrak{L}$ and extend this to a relation $\prec_2$ on $\mathfrak{L} \times W \times \mathcal{A}$ defined as follows: $(s, v, \mu) \prec_2 (t, w, \nu)$ if $s \prec t$. It is easy to see that $\prec_2$ is a well-founded order on $\mathfrak{L} \times W \times \mathcal{A}$. The minimal elements of $\mathfrak{L} \times W \times \mathcal{A}$ are triples of the form (t, w, ν), where t is either a variable or a constant. Also needed are two partitions: $\{\mathfrak{L}_\alpha \times W \times \mathcal{A}\}_{\alpha\in\mathfrak{S}}$ of $\mathfrak{L} \times W \times \mathcal{A}$ and $\{\mathfrak{D}_\alpha\}_{\alpha\in\mathfrak{S}}$ of $\mathfrak{D}$. Then it needs to be checked that the two conditions concerning consistency in Proposition B.4.4 are satisfied. First, each minimal element is consistent since each denotation of a variable or constant of type α is defined to be an element of $\mathfrak{D}_\alpha$. Second, by considering the last four cases in Definition B.2.10, it is clear that the rule defining T has the property that if (s, v, μ) is consistent, for each $(s, v, \mu) \prec_2 (t, w, \nu)$, then (t, w, ν) is consistent. Thus, by Proposition B.4.4, the function

T exists, is unique, and satisfies the condition $T(\mathfrak{L}_\alpha \times W \times \mathcal{A}) \subseteq \mathfrak{D}_\alpha$, for all $\alpha \in \mathfrak{S}$. The latter condition states exactly that if t is a term of type α, $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, and w a world in W, then $\mathcal{V}(t, I, w, \nu) \in \mathcal{D}_\alpha$.

The next three results establish (mostly) familiar properties of the connectives and quantifiers, but in the more general setting of biterms rather than formulas.

Proposition B.2.1. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, φ and ψ biterms of type $\alpha_1 \to \cdots \to \alpha_n \to o$, $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, and t a term of type α_1. Then the following hold.*

1. $\mathcal{V}(\varphi \wedge \psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$
   iff $\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$ *and* $\mathcal{V}(\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$.

2. $\mathcal{V}(\varphi \vee \psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$
   iff $\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$ *or* $\mathcal{V}(\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$.

3. $\mathcal{V}(\varphi \longrightarrow \psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$
   iff $\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{F}$ *or* $\mathcal{V}(\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$.

4. $\mathcal{V}(\neg\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$ *iff* $\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{F}$.

5. $\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$ *iff* $\mathcal{V}(\neg\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{F}$.

6. $\mathcal{V}(\neg\neg\varphi, I, w, \nu) = \mathcal{V}(\varphi, I, w, \nu)$.

7. $\mathcal{V}(\neg(\varphi \wedge \psi), I, w, \nu) = \mathcal{V}(\neg\varphi \vee \neg\psi, I, w, \nu)$.

8. $\mathcal{V}(\neg(\varphi \vee \psi), I, w, \nu) = \mathcal{V}(\neg\varphi \wedge \neg\psi, I, w, \nu)$.

9. $\mathcal{V}((\neg\varphi\ t), I, w, \nu) = \mathcal{V}(\neg(\varphi\ t), I, w, \nu)$.

10. $\mathcal{V}(\lambda x.\neg\varphi, I, w, \nu) = \mathcal{V}(\neg\lambda x.\varphi, I, w, \nu)$.

11. $\mathcal{V}(\lambda x.\varphi \wedge \lambda x.\psi, I, w, \nu) = \mathcal{V}(\lambda x.(\varphi \wedge \psi), I, w, \nu)$.

12. $\mathcal{V}(\lambda x.\varphi \vee \lambda x.\psi, I, w, \nu) = \mathcal{V}(\lambda x.(\varphi \vee \psi), I, w, \nu)$.

13. $\mathcal{V}(\lambda x.\varphi \longrightarrow \lambda x.\psi, I, w, \nu) = \mathcal{V}(\lambda x.(\varphi \longrightarrow \psi), I, w, \nu)$.

Proof. 1.

$$
\begin{aligned}
&\mathcal{V}(\varphi \wedge \psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}\\
\text{iff}\ \ &\mathcal{V}((\wedge\ \varphi), I, w, \nu)(\mathcal{V}(\psi, I, w, \nu))\ d_1\ \ldots\ d_n = \mathsf{T}\\
\text{iff}\ \ &(\mathcal{V}(\wedge, I, w, \nu)(\mathcal{V}(\varphi, I, w, \nu)))(\mathcal{V}(\psi, I, w, \nu))\ d_1\ \ldots\ d_n = \mathsf{T}\\
\text{iff}\ \ &(V(\wedge, w)(\mathcal{V}(\varphi, I, w, \nu)))(\mathcal{V}(\psi, I, w, \nu))\ d_1\ \ldots\ d_n = \mathsf{T}\\
\text{iff}\ \ &\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \text{ and } \mathcal{V}(\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.
\end{aligned}
$$

2.

$$
\begin{aligned}
&\mathcal{V}(\varphi \vee \psi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T} \\
\text{iff }\ &\mathcal{V}((\vee\, \varphi), I, w, \nu)(\mathcal{V}(\psi, I, w, \nu))\ d_1 \ldots d_n = \mathsf{T} \\
\text{iff }\ &(\mathcal{V}(\vee, I, w, \nu)(\mathcal{V}(\varphi, I, w, \nu)))(\mathcal{V}(\psi, I, w, \nu))\ d_1 \ldots d_n = \mathsf{T} \\
\text{iff }\ &(V(\vee, w)(\mathcal{V}(\varphi, I, w, \nu)))(\mathcal{V}(\psi, I, w, \nu))\ d_1 \ldots d_n = \mathsf{T} \\
\text{iff }\ &\mathcal{V}(\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T} \text{ or } \mathcal{V}(\psi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T}.
\end{aligned}
$$

3, Similar to the proof of Part 2.

4.

$$
\begin{aligned}
&\mathcal{V}(\neg\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T} \\
\text{iff }\ &\mathcal{V}(\neg, I, w, \nu)(\mathcal{V}(\varphi, I, w, \nu))\ d_1 \ldots d_n = \mathsf{T} \\
\text{iff }\ &V(\neg, w)(\mathcal{V}(\varphi, I, w, \nu))\ d_1 \ldots d_n = \mathsf{T} \\
\text{iff }\ &\mathcal{V}(\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{F}.
\end{aligned}
$$

5. Similar to the proof of Part 4.

6.

$$
\begin{aligned}
&\mathcal{V}(\neg\neg\varphi, I, w, \nu) \\
&= \mathcal{V}(\neg, I, w, \nu)(\mathcal{V}(\neg\varphi, I, w, \nu)) \\
&= V(\neg, w)(\mathcal{V}(\neg\varphi, I, w, \nu)) \\
&= V(\neg, w)(V(\neg, w)(\mathcal{V}(\varphi, I, w, \nu))) \\
&= \mathcal{V}(\varphi, I, w, \nu). \qquad \text{[Definition of } V(\neg, w)\text{]}.
\end{aligned}
$$

7.

$$
\begin{aligned}
&\mathcal{V}(\neg(\varphi \wedge \psi), I, w, \nu)\ d_1 \ldots d_n \\
&= V(\neg, w)(\mathcal{V}(\varphi \wedge \psi, I, w, \nu))\ d_1 \ldots d_n \\
&= \begin{cases} \mathsf{T} & \text{if } \mathcal{V}(\varphi \wedge \psi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{F} \\ \mathsf{F} & \text{otherwise} \end{cases} \\
&= \begin{cases} \mathsf{T} & \text{if } (V(\wedge, w)(\mathcal{V}(\varphi, I, w, \nu)))(\mathcal{V}(\psi, I, w, \nu))\ d_1 \ldots d_n = \mathsf{F} \\ \mathsf{F} & \text{otherwise} \end{cases} \\
&= \begin{cases} \mathsf{T} & \text{if } \mathcal{V}(\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{F} \text{ or } \mathcal{V}(\psi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{F} \\ \mathsf{F} & \text{otherwise} \end{cases} \\
&= \begin{cases} \mathsf{T} & \text{if } V(\neg, w)(\mathcal{V}(\varphi, I, w, \nu))\ d_1 \ldots d_n = \mathsf{T} \text{ or} \\ & \quad V(\neg, w)(\mathcal{V}(\psi, I, w, \nu))\ d_1 \ldots d_n = \mathsf{T} \\ \mathsf{F} & \text{otherwise} \end{cases} \\
&= (V(\vee, w)(V(\neg, w)(\mathcal{V}(\varphi, I, w, \nu))))(V(\neg, w)(\mathcal{V}(\psi, I, w, \nu)))\ d_1 \ldots d_n \\
&= (V(\vee, w)(\mathcal{V}(\neg\varphi, I, w, \nu)))(\mathcal{V}(\neg\psi, I, w, \nu))\ d_1 \ldots d_n \\
&= \mathcal{V}(\neg\varphi \vee \neg\psi, I, w, \nu)\ d_1 \ldots d_n.
\end{aligned}
$$

Hence $\mathcal{V}(\neg(\varphi \wedge \psi), I, w, \nu) = \mathcal{V}(\neg\varphi \vee \neg\psi, I, w, \nu)$.

8. Similar to the proof of Part 7.

9.

$$
\begin{aligned}
&\mathcal{V}((\neg\varphi\ t), I, w, \nu)\\
&= \mathcal{V}(\neg\varphi, I, w, \nu)(\mathcal{V}(t, I, w, \nu))\\
&= (\mathcal{V}(\neg, I, w, \nu)(\mathcal{V}(\varphi, I, w, \nu)))(\mathcal{V}(t, I, w, \nu))\\
&= (V(\neg, w)(\mathcal{V}(\varphi, I, w, \nu)))(\mathcal{V}(t, I, w, \nu))\\
&= V(\neg, w)(\mathcal{V}(\varphi, I, w, \nu)(\mathcal{V}(t, I, w, \nu))) \qquad \text{[Definition of } V(\neg, w)\text{]}\\
&= V(\neg, w)(\mathcal{V}((\varphi\ t), I, w, \nu))\\
&= \mathcal{V}(\neg(\varphi\ t), I, w, \nu).
\end{aligned}
$$

10.

$$
\begin{aligned}
&\mathcal{V}(\lambda x.\neg\varphi, I, w, \nu)\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(\neg\varphi, I, w, \nu'),\\
&\quad \text{where } x \text{ has type } \alpha \text{ and } \nu' \text{ is } \nu \text{ except } \nu'(x) = d\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } V(\neg, w)(\mathcal{V}(\varphi, I, w, \nu'))\\
&= V(\neg, w)(\text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(\varphi, I, w, \nu'))\\
&= V(\neg, w)(\mathcal{V}(\lambda x.\varphi, I.w, \nu))\\
&= \mathcal{V}(\neg\lambda x.\varphi, I, w, \nu).
\end{aligned}
$$

11.

$$
\begin{aligned}
&\mathcal{V}(\lambda x.\varphi \wedge \lambda x.\psi, I, w, \nu)\\
&= \mathcal{V}((\wedge\ \lambda x.\varphi), I, w, \nu)(\mathcal{V}(\lambda x.\psi, I, w, \nu))\\
&= (V(\wedge, w)(\mathcal{V}(\lambda x.\varphi, I, w, \nu)))(\mathcal{V}(\lambda x.\psi, I, w, \nu))\\
&= (V(\wedge, w)(\text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(\varphi, I, w, \nu')))\\
&\qquad\qquad (\text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(\psi, I, w, \nu')),\\
&\quad \text{where } x \text{ has type } \alpha \text{ and } \nu' \text{ is } \nu \text{ except } \nu'(x) = d\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is}\\
&\quad (V(\wedge, w)(\mathcal{V}(\varphi, I, w, \nu')))(\mathcal{V}(\psi, I, w, \nu'))\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(\varphi \wedge \psi, I, w, \nu')\\
&= \mathcal{V}(\lambda x.(\varphi \wedge \psi), I, w, \nu).
\end{aligned}
$$

12. Similar to the proof of Part 11.

13. Similar to the proof of Part 11. □

Proposition B.2.2. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, and φ a biterm having type $\alpha_1 \to \cdots \to \alpha_n \to o$. Then the following hold.*

1. $\mathcal{V}((\Sigma\ \varphi), I, w, \nu) = \mathsf{T}$
   iff, for some $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, $\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$.

2. $\mathcal{V}((\Pi\ \varphi), I, w, \nu) = \mathsf{T}$
iff, for each $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, $\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$.

Proof. 1.

$$\begin{aligned}
&\mathcal{V}((\Sigma\ \varphi), I, w, \nu) = \mathsf{T} \\
\text{iff}\ &V(\Sigma, w)(\mathcal{V}(\varphi, I, w, \nu)) = \mathsf{T} \\
\text{iff}\ &\mathcal{V}(\varphi, I, w, \nu) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \\
\text{iff}\ &\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \text{ for some } d_i \in \mathcal{D}_{\alpha_i}\ (i = 1, \ldots, n).
\end{aligned}$$

2.

$$\begin{aligned}
&\mathcal{V}((\Pi\ \varphi), I, w, \nu) = \mathsf{T} \\
\text{iff}\ &V(\Pi, w)(\mathcal{V}(\varphi, I, w, \nu)) = \mathsf{T} \\
\text{iff}\ &\mathcal{V}(\varphi, I, w, \nu) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \\
\text{iff}\ &\mathcal{V}(\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \text{ for each } d_i \in \mathcal{D}_{\alpha_i}\ (i = 1, \ldots, n).
\end{aligned}$$

□

Proposition B.2.3. *Let* $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ *be an interpretation,* ν *a variable assignment with respect to* $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, *and* φ *a biterm. Then the following hold.*

1. $\mathcal{V}(\neg(\Sigma\ \varphi), I, w, \nu) = \mathcal{V}((\Pi\ \neg\varphi), I, w, \nu)$.

2. $\mathcal{V}(\neg(\Pi\ \varphi), I, w, \nu) = \mathcal{V}((\Sigma\ \neg\varphi), I, w, \nu)$.

Proof. 1.

$$\begin{aligned}
&\mathcal{V}(\neg(\Sigma\ \varphi), I, w, \nu) \\
&= V(\neg, w)(\mathcal{V}((\Sigma\ \varphi), I, w, \nu)) \\
&= V(\neg, w)(V(\Sigma, w)(\mathcal{V}(\varphi, I, w, v))) \\
&= V(\neg, w)(\text{if } \mathcal{V}(\varphi, I, w, v) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \text{ then } \mathsf{T} \text{ else } \mathsf{F}) \\
&= \text{if } \mathcal{V}(\varphi, I, w, v) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \text{ then } \mathsf{F} \text{ else } \mathsf{T} \\
&= \text{if } \mathcal{V}(\varphi, I, w, v) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \text{ then } \mathsf{T} \text{ else } \mathsf{F} \\
&= \text{if } V(\neg, w)(\mathcal{V}(\varphi, I, w, v)) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \text{ then } \mathsf{T} \text{ else } \mathsf{F} \\
&= \text{if } \mathcal{V}(\neg\varphi, I, w, v) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \text{ then } \mathsf{T} \text{ else } \mathsf{F} \\
&= V(\Pi, w)(\mathcal{V}(\neg\varphi, I, w, v)) \\
&= \mathcal{V}((\Pi\ \neg\varphi), I, w, \nu).
\end{aligned}$$

2.

$$\begin{aligned}
&\mathcal{V}(\neg(\Pi\ \varphi), I, w, \nu)\\
&= V(\neg, w)(\mathcal{V}((\Pi\ \varphi), I, w, \nu))\\
&= V(\neg, w)(V(\Pi, w)(\mathcal{V}(\varphi, I, w, v)))\\
&= V(\neg, w)(\text{if } \mathcal{V}(\varphi, I, w, v) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \text{ then } \mathsf{T} \text{ else } \mathsf{F})\\
&= \text{if } \mathcal{V}(\varphi, I, w, v) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \text{ then } \mathsf{F} \text{ else } \mathsf{T}\\
&= \text{if } \mathcal{V}(\varphi, I, w, v) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \text{ then } \mathsf{T} \text{ else } \mathsf{F}\\
&= \text{if } V(\neg, w)(\mathcal{V}(\varphi, I, w, v)) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \text{ then } \mathsf{T} \text{ else } \mathsf{F}\\
&= \text{if } \mathcal{V}(\neg\varphi, I, w, v) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \text{ then } \mathsf{T} \text{ else } \mathsf{F}\\
&= V(\Sigma, w)(\mathcal{V}(\neg\varphi, I, w, v))\\
&= \mathcal{V}((\Sigma\ \neg\varphi), I, w, \nu).
\end{aligned}$$

□

Definition B.2.11. The variable assignments ν and ν' are said to be *x-variants* if they agree on all variables except possibly the variable x.

Proposition B.2.4. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}},\ V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, and φ a biterm of rank n. Then the following hold.*

1. $\mathcal{V}(\lambda x.\varphi, I, w, \nu) = \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}$ *iff, for every x-variant ν' of ν, $\mathcal{V}(\varphi, I, w, \nu') = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}$.*
2. $\mathcal{V}(\lambda x.\varphi, I, w, \nu) = \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}$ *iff, for every x-variant ν' of ν, $\mathcal{V}(\varphi, I, w, \nu') = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}$.*
3. $\mathcal{V}(\forall x.\varphi, I, w, \nu) = \mathsf{T}$ *iff, for every x-variant ν' of ν, $\mathcal{V}(\varphi, I, w, \nu') = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}$.*
4. $\mathcal{V}(\exists x.\varphi, I, w, \nu) = \mathsf{T}$ *iff, for some x-variant ν' of ν, $\mathcal{V}(\varphi, I, w, \nu') \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}$.*
5. $\mathcal{V}(\neg\exists x.\varphi, I, w, \nu) = \mathcal{V}(\forall x.\neg\varphi, I, w, \nu)$.
6. $\mathcal{V}(\neg\forall x.\varphi, I, w, \nu) = \mathcal{V}(\exists x.\neg\varphi, I, w, \nu)$.
7. $\mathcal{V}(\forall x.\varphi, I, w, \nu) = \mathcal{V}(\neg\exists x.\neg\varphi, I, w, \nu)$.
8. $\mathcal{V}(\exists x.\varphi, I, w, \nu) = \mathcal{V}(\neg\forall x.\neg\varphi, I, w, \nu)$.

Proof. 1.

$$\begin{aligned}
&\mathcal{V}(\lambda x.\varphi, I, w, \nu) = \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}\\
&\text{iff } \mathcal{V}(\varphi, I, w, \nu') = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}, \text{ for every } x\text{-variant } \nu' \text{ of } \nu. \quad [\text{Semantics of } \lambda x.\varphi]
\end{aligned}$$

2.

$$\begin{aligned}
&\mathcal{V}(\lambda x.\varphi, I, w, \nu) = \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}\\
&\text{iff } \mathcal{V}(\varphi, I, w, \nu') = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}, \text{ for every } x\text{-variant } \nu' \text{ of } \nu. \quad [\text{Semantics of } \lambda x.\varphi]
\end{aligned}$$

3.

$$\begin{aligned}
&\mathcal{V}(\forall x.\varphi, I, w, \nu) = \mathsf{T} \\
\text{iff } &V(\Pi, w)(\mathcal{V}(\lambda x.\varphi, I, w, \nu)) = \mathsf{T} \\
\text{iff } &\mathcal{V}(\lambda x.\varphi, I, w, \nu) = \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \\
\text{iff } &\mathcal{V}(\varphi, I, w, \nu') = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}, \text{ for every } x\text{-variant } \nu' \text{ of } \nu. \qquad \text{[Part 1]}
\end{aligned}$$

4.

$$\begin{aligned}
&\mathcal{V}(\exists x.\varphi, I, w, \nu) = \mathsf{T} \\
\text{iff } &V(\Sigma, w)(\mathcal{V}(\lambda x.\varphi, I, w, \nu)) = \mathsf{T} \\
\text{iff } &\mathcal{V}(\lambda x.\varphi, I, w, \nu) \neq \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F} \\
\text{iff } &\mathcal{V}(\varphi, I, w, \nu') \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}, \text{ for some } x\text{-variant } \nu' \text{ of } \nu. \qquad \text{[Part 2]}
\end{aligned}$$

5.

$$\begin{aligned}
&\mathcal{V}(\neg\exists x.\varphi, I, w, \nu) \\
&= \mathcal{V}(\neg(\Sigma\ \lambda x.\varphi), I, w, \nu) \\
&= \mathcal{V}((\Pi\ \neg\lambda x.\varphi), I, w, \nu) && \text{[Part 1 of Proposition B.2.3]} \\
&= V(\Pi, w)(\mathcal{V}(\neg\lambda x.\varphi, I, w, \nu)) \\
&= V(\Pi, w)(\mathcal{V}(\lambda x.\neg\varphi, I, w, \nu)) && \text{[Part 10 of Proposition B.2.1]} \\
&= \mathcal{V}((\Pi\ \lambda x.\neg\varphi), I, w, \nu) \\
&= \mathcal{V}(\forall x.\neg\varphi, I, w, \nu).
\end{aligned}$$

6.

$$\begin{aligned}
&\mathcal{V}(\neg\forall x.\varphi, I, w, \nu) \\
&= \mathcal{V}(\neg(\Pi\ \lambda x.\varphi), I, w, \nu) \\
&= \mathcal{V}((\Sigma\ \neg\lambda x.\varphi), I, w, \nu) && \text{[Part 2 of Proposition B.2.3]} \\
&= V(\Sigma, w)(\mathcal{V}(\neg\lambda x.\varphi, I, w, \nu)) \\
&= V(\Sigma, w)(\mathcal{V}(\lambda x.\neg\varphi, I, w, \nu)) && \text{[Part 10 of Proposition B.2.1]} \\
&= \mathcal{V}((\Sigma\ \lambda x.\neg\varphi), I, w, \nu) \\
&= \mathcal{V}(\exists x.\neg\varphi, I, w, \nu).
\end{aligned}$$

7. This part follows from Part 6 of Proposition B.2.1 and Part 5.
8. This part follows from Part 6 of Proposition B.2.1 and Part 6.

□

Now come some useful properties of modalities.

Proposition B.2.5. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, and t a term. Then*

$$\mathcal{V}(\Box_{j_1}\cdots\Box_{j_n}t, I, w, \nu) = \mathcal{M}(\{\mathcal{V}(t, I, w', \nu)\}_{w'\in W'}),$$

where $W' = \{w' : w\, R_{j_1} \circ \cdots \circ R_{j_n}\, w'\}$.

Proof. The proof is by induction on the length of the sequence of modalities. When $n = 1$, the result follows immediately from the definition of denotation for modal terms. Assume now that the result holds for sequences of modalities of length n. Then

$$\begin{aligned}
&\mathcal{V}(\Box_{j_1} \cdots \Box_{j_{n+1}} t, I, w, \nu) \\
&= \mathcal{M}(\{\mathcal{V}(\Box_{j_2} \cdots \Box_{j_{n+1}} t, I, w'', \nu)\}_{w'' \in W''}), \text{ where } W'' = \{w'' : w\, R_{j_1}\, w''\} \\
&= \mathcal{M}(\{\mathcal{M}(\{\mathcal{V}(t, I, w', \nu)\}_{w' \in W'})\}_{w'' \in W''}), \\
&\qquad \text{where } W'' = \{w'' : w\, R_{j_1}\, w''\} \text{ and } W' = \{w' : w''\, R_{j_2} \circ \cdots \circ R_{j_{n+1}}\, w'\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[Induction hypothesis]} \\
&= \mathcal{M}(\{\mathcal{V}(t, I, w', \nu)\}_{w' \in W'}), \text{ where } W' = \{w' : w\, R_{j_1} \circ \cdots \circ R_{j_{n+1}}\, w'\}
\end{aligned}$$

Hence the result. □

Proposition B.2.5 shows that $\Box_{j_1} \cdots \Box_{j_n}$ can be treated as a kind of (composite) modality with accessibility relation $R_{j_1} \circ \cdots \circ R_{j_n}$.

Proposition B.2.6. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}$, $w \in W$, $i \in \{1, \ldots, m\}$, $d_j \in \mathcal{D}_{\alpha_j} (j = 1, \ldots, n)$, and φ and ψ biterms having type $\alpha_1 \to \cdots \to \alpha_n \to o$. Then the following hold.*

1. $\mathcal{V}(\Box_i \varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$ *iff, for each w' such that $w\, R_i\, w'$, $\mathcal{V}(\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$.*
2. $\mathcal{V}(\Diamond_i \varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$ *iff, for some w' such that $w\, R_i\, w'$, $\mathcal{V}(\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$.*
3. $\mathcal{V}(\Box_i \varphi, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$ *iff, for each w' such that $w\, R_i\, w'$, $\mathcal{V}(\varphi, I, w', \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$.*
4. $\mathcal{V}(\Diamond_i \varphi, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$ *iff, for some w' such that $w\, R_i\, w'$, $\mathcal{V}(\varphi, I, w', \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$.*
5. $\mathcal{V}(\Box_i(\varphi \wedge \psi), I, w, \nu) = \mathcal{V}(\Box_i \varphi \wedge \Box_i \psi, I, w, \nu)$.
6. $\mathcal{V}(\Diamond_i(\varphi \vee \psi), I, w, \nu) = \mathcal{V}(\Diamond_i \varphi \vee \Diamond_i \psi, I, w, \nu)$.

Proof. 1.

$$\begin{aligned}
&\mathcal{V}(\Box_i \varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \\
\text{iff } &\mathcal{M}(\{\mathcal{V}(\varphi, I, w', \nu)\}_{w' \in W'})\ d_1\ \ldots\ d_n = \mathsf{T} \\
\text{iff } &\mathcal{V}(\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \text{ for each } w' \text{ such that } w\, R_i\, w'.
\end{aligned}$$

2.

$$\begin{aligned}
&\mathcal{V}(\Diamond_i \varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \\
\text{iff } &\mathcal{V}(\neg\Box_i\neg\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \\
\text{iff } &\mathcal{V}(\Box_i\neg\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{F} \\
\text{iff } &\mathcal{M}(\{\mathcal{V}(\neg\varphi, I, w', \nu)\}_{w' \in W'})\ d_1\ \ldots\ d_n = \mathsf{F} \\
\text{iff } &\mathcal{V}(\neg\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{F}, \text{ for some } w' \text{ such that } w\, R_i\, w'. \\
\text{iff } &\mathcal{V}(\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \text{ for some } w' \text{ such that } w\, R_i\, w'.
\end{aligned}$$

3.

$$\begin{aligned}
&\mathcal{V}(\Box_i\varphi, I, w, \nu) \;=\; \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}\\
\text{iff }\; &\mathcal{M}(\{\mathcal{V}(\varphi, I, w', \nu)\}_{w'\in W'}) \;=\; \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}\\
\text{iff }\; &\mathcal{V}(\varphi, I, w', \nu) \;=\; \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}, \text{ for each } w' \text{ such that } w\, R_i\, w'.
\end{aligned}$$

4.

$$\begin{aligned}
&\mathcal{V}(\Diamond_i\varphi, I, w, \nu) \;=\; \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}\\
\text{iff }\; &\mathcal{V}(\neg\Box_i\neg\varphi, I, w, \nu) \;=\; \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}\\
\text{iff }\; &\mathcal{V}(\Box_i\neg\varphi, I, w, \nu) \;=\; \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}\\
\text{iff }\; &\text{for some } w' \text{ such that } w\, R_i\, w',\ \mathcal{V}(\neg\varphi, I, w', \nu) \;=\; \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}\\
\text{iff }\; &\text{for some } w' \text{ such that } w\, R_i\, w',\ \mathcal{V}(\varphi, I, w', \nu) \;=\; \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}.
\end{aligned}$$

5.

$$\begin{aligned}
&\mathcal{V}(\Box_i(\varphi\wedge\psi), I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}\\
\text{iff }\; &\mathcal{V}(\varphi\wedge\psi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \text{ for each } w' \text{ such that } w\, R_i\, w' && \text{[Part 1]}\\
\text{iff }\; &\mathcal{V}(\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \text{ and } \mathcal{V}(\psi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T},\\
&\quad \text{for each } w' \text{ such that } w\, R_i\, w'\\
\text{iff }\; &\mathcal{V}(\Box_i\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \text{ and } \mathcal{V}(\Box_i\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} && \text{[Part 1]}\\
\text{iff }\; &\mathcal{V}(\Box_i\varphi\wedge\Box_i\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.
\end{aligned}$$

Hence $\mathcal{V}(\Box_i(\varphi\wedge\psi), I, w, \nu) = \mathcal{V}(\Box_i\varphi\wedge\Box_i\psi, I, w, \nu)$

6.

$$\begin{aligned}
&\mathcal{V}(\Diamond_i(\varphi\vee\psi), I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}\\
\text{iff }\; &\mathcal{V}(\varphi\vee\psi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \text{ for some } w' \text{ such that } w\, R_i\, w' && \text{[Part 2]}\\
\text{iff }\; &\mathcal{V}(\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \text{ or } \mathcal{V}(\psi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T},\\
&\quad \text{for some } w' \text{ such that } w\, R_i\, w'\\
\text{iff }\; &\mathcal{V}(\Diamond_i\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \text{ or } \mathcal{V}(\Diamond_i\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} && \text{[Part 2]}\\
\text{iff }\; &\mathcal{V}(\Diamond_i\varphi\vee\Diamond_i\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.
\end{aligned}$$

Hence $\mathcal{V}(\Diamond_i(\varphi\vee\psi), I, w, \nu) = \mathcal{V}(\Diamond_i\varphi\vee\Diamond_i\psi, I, w, \nu)$

□

Proposition B.2.7. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, t be a term of type α having rank 0 and d a non-default element of $\mathcal{D}_\alpha$. Then $\mathcal{V}(\Box_{j_1}\cdots\Box_{j_n}t, I, w, \nu) = d$ iff, for each w' such that $w\, R_{j_1}\circ\cdots\circ R_{j_n}\, w'$, $\mathcal{V}(t, I, w', \nu) = d$.*

Proof. The proof is by induction on the length of the sequence of modalities. If $n = 0$, the result is obvious.

Suppose now the result holds for sequences of modalities of length n. Then

$$\begin{aligned}
&\mathcal{V}(\Box_{j_1}\cdots\Box_{j_{n+1}}t, I, w, \nu) = d\\
\text{iff}\ \ &\text{for each } w'' \text{ such that } w\,R_{j_1}\,w'', \mathcal{V}(\Box_{j_2}\cdots\Box_{j_{n+1}}t, I, w'', \nu) = d\\
\text{iff}\ \ &\text{for each } w'' \text{ such that } w\,R_{j_1}\,w'', \text{ for each } w' \text{ such that}\\
&w''\,R_{j_2}\circ\cdots\circ R_{j_{n+1}}\,w', \mathcal{V}(t, I, w', \nu) = d \qquad \text{[Induction hypothesis]}\\
\text{iff}\ \ &\text{for each } w' \text{ such that } w\,R_{j_1}\circ\cdots\circ R_{j_{n+1}}\,w',\ \mathcal{V}(t, I, w', \nu) = d.
\end{aligned}$$

□

The next result shows that $\Box_i$ and Π can be 'switched', as can $\Diamond_i$ and Σ.

Proposition B.2.8. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, φ a biterm, and $i \in \{1, \ldots, m\}$. Then the following hold.*

1. $\mathcal{V}(\Box_i(\Pi\ \varphi), I, w, \nu) = \mathcal{V}((\Pi\ \Box_i\varphi), I, w, \nu)$.
2. $\mathcal{V}(\Diamond_i\Sigma\ \varphi), I, w, \nu) = \mathcal{V}((\Sigma\ \Diamond_i\varphi), I, w, \nu)$.

Proof. Suppose that φ has type $\alpha_1 \to \cdots \to \alpha_n \to o$.

1.

$$\begin{aligned}
&\mathcal{V}(\Box_i(\Pi\ \varphi), I, w, \nu) = \mathsf{T}\\
\text{iff}\ \ &\mathcal{V}((\Pi\ \varphi), I, w', \nu) = \mathsf{T}, \text{ for each } w' \text{ such that } w\,R_i\,w' && \text{[Prop. B.2.6]}\\
\text{iff}\ \ &\mathcal{V}(\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T},\\
&\text{for each } w' \text{ such that } w\,R_i\,w' \text{ and each } d_j \in \mathcal{D}_{\alpha_j} (j = 1, \ldots, n) && \text{[Prop. B.2.2]}\\
\text{iff}\ \ &\mathcal{V}(\Box_i\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \text{ for each } d_j \in \mathcal{D}_{\alpha_j} (j = 1, \ldots, n) && \text{[Prop. B.2.6]}\\
\text{iff}\ \ &\mathcal{V}((\Pi\ \Box_i\varphi), I, w, \nu) = \mathsf{T}. && \text{[Prop. B.2.2]}
\end{aligned}$$

Hence $\mathcal{V}(\Box_i(\Pi\ \varphi), I, w, \nu) = \mathcal{V}((\Pi\ \Box_i\varphi), I, w, \nu)$.

2. The proof is similar to Part 1.

□

Semantic versions of the Barcan and converse Barcan formulas are now established, in the biterm setting. (So they are called the Barcan biterm and converse Barcan biterm.)

Proposition B.2.9. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, φ a biterm, and $i \in \{1, \ldots, m\}$. Then the following hold.*

1. $\mathcal{V}(\Box_i\forall x.\varphi, I, w, \nu) = \mathcal{V}(\forall x.\Box_i\varphi, I, w, \nu)$.
2. $\mathcal{V}(\Diamond_i\exists x.\varphi, I, w, \nu) = \mathcal{V}(\exists x.\Diamond_i\varphi, I, w, \nu)$.

Proof. Suppose that φ has type $\alpha_1 \to \cdots \to \alpha_n \to o$ and x has type α.

1.

$$
\begin{aligned}
&\mathcal{V}(\Box_i \forall x.\varphi, I, w, \nu) \\
&= \mathcal{M}(\{\mathcal{V}(\forall x.\varphi, I, w', \nu)\}_{w' \in W'}) \\
&= \mathcal{M}(\{\mathcal{V}((\Pi\ \lambda x.\varphi), I, w', \nu)\}_{w' \in W'}) \\
&= \mathcal{M}(\{V(\Pi, w')(\mathcal{V}(\lambda x.\varphi, I, w', \nu))\}_{w' \in W'}) \\
&= \mathcal{M}(\{\text{if } \mathcal{V}(\lambda x.\varphi, I, w', \nu) = \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \text{ then } \mathsf{T} \text{ else } \mathsf{F}\}_{w' \in W'}), \\
&\quad \text{where } \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} : \mathcal{D}_\alpha \to \mathcal{D}_{\alpha_1} \to \cdots \to \mathcal{D}_{\alpha_n} \to \mathcal{D}_o \\
&= \mathcal{M}(\{\text{if } \mathcal{V}(\varphi, I, w', \nu') = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}, \text{ for all } d \in \mathcal{D}_\alpha, \text{ then } \mathsf{T} \text{ else } \mathsf{F}\}_{w' \in W'}), \\
&\quad \text{where } \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} : \mathcal{D}_{\alpha_1} \to \cdots \to \mathcal{D}_{\alpha_n} \to \mathcal{D}_o \text{ and } \nu' \text{ is } \nu \text{ except } \nu'(x) = d \\
&= \text{if } \mathcal{M}(\{\mathcal{V}(\varphi, I, w', \nu')\}_{w' \in W'}) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}, \text{ for all } d \in \mathcal{D}_\alpha, \text{ then } \mathsf{T} \text{ else } \mathsf{F} \\
&= \text{if } \mathcal{V}(\Box_i \varphi, I, w, \nu') = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}, \text{ for all } d \in \mathcal{D}_\alpha, \text{ then } \mathsf{T} \text{ else } \mathsf{F} \\
&= \text{if } \mathcal{V}(\lambda x.\Box_i \varphi, I, w, \nu) = \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} \text{ then } \mathsf{T} \text{ else } \mathsf{F}, \\
&\quad \text{where } \lambda x_0.\lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T} : \mathcal{D}_\alpha \to \mathcal{D}_{\alpha_1} \to \cdots \to \mathcal{D}_{\alpha_n} \to \mathcal{D}_o \\
&= V(\Pi, w)(\mathcal{V}(\lambda x.\Box_i \varphi, I, w, \nu)) \\
&= \mathcal{V}((\Pi\ \lambda x.\Box_i \varphi), I, w, \nu) \\
&= \mathcal{V}(\forall x.\Box_i \varphi, I, w, \nu).
\end{aligned}
$$

2.

$$
\begin{aligned}
&\mathcal{V}(\Diamond_i \exists x.\varphi, I, w, \nu) \\
&= \mathcal{V}(\neg\Box_i\neg\exists x.\varphi, I, w, \nu) \\
&= V(\neg, w)(\mathcal{V}(\Box_i\neg\exists x.\varphi, I, w, \nu)) \\
&= V(\neg, w)(\mathcal{M}(\{\mathcal{V}(\neg\exists x.\varphi, I, w', \nu)\}_{w' \in W'})) \\
&= V(\neg, w)(\mathcal{M}(\{\mathcal{V}(\forall x.\neg\varphi, I, w', \nu)\}_{w' \in W'})) && \text{[Part 5 of Proposition B.2.4]} \\
&= V(\neg, w)(\mathcal{V}(\Box_i \forall x.\neg\varphi, I, w, \nu)) \\
&= V(\neg, w)(\mathcal{V}(\forall x.\Box_i\neg\varphi, I, w, \nu)) && \text{[Part 1]} \\
&= \mathcal{V}(\neg\forall x.\Box_i\neg\varphi, I, w, \nu) \\
&= \mathcal{V}(\exists x.\neg\Box_i\neg\varphi, I, w, \nu) && \text{[Part 6 of Proposition B.2.4]} \\
&= \mathcal{V}(\exists x.\Diamond_i \varphi, I, w, \nu).
\end{aligned}
$$

□

Proposition B.2.10. *Let s and t be terms of the same type, I an interpretation, w a world in I, and ν a variable assignment. Then $\mathcal{V}(s = t, I, w, \nu) = \mathsf{T}$ iff $\mathcal{V}(s, I, w, \nu) = \mathcal{V}(t, I, w, \nu)$.*

Proof.

$$
\begin{aligned}
&\mathcal{V}(s = t, I, w, \nu) = \mathsf{T} \\
&\text{iff } \mathcal{V}((= s), I, w, \nu)\ (\mathcal{V}(t, I, w, \nu)) = \mathsf{T} \\
&\text{iff } (\mathcal{V}(=, I, w, \nu)\ (\mathcal{V}(s, I, w, \nu)))\ (\mathcal{V}(t, I, w, \nu)) = \mathsf{T} \\
&\text{iff } \mathcal{V}(s, I, w, \nu) = \mathcal{V}(t, I, w, \nu).
\end{aligned}
$$

□

The bi-implication connective $\longleftrightarrow$: $\alpha \to \alpha \to \alpha$, where α is a biterm type, can be introduced by the abbreviation $\varphi \longleftrightarrow \psi \triangleq (\varphi \longrightarrow \psi) \wedge (\psi \longrightarrow \varphi)$. The next result implies that, applied to formulas, $\longleftrightarrow$ is just equality.

Proposition B.2.11. *Let φ and ψ be biterms of type $\alpha_1 \to \cdots \to \alpha_n \to o$, I an interpretation, w a world in I, and ν a variable assignment. Then $\mathcal{V}(\varphi \longleftrightarrow \psi, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$ iff $\mathcal{V}(\varphi = \psi, I, w, \nu) = \mathsf{T}$.*

Proof.

$\mathcal{V}((\varphi \longrightarrow \psi) \wedge (\psi \longrightarrow \varphi), I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$
iff $(V(\wedge, w)(\mathcal{V}(\varphi \longrightarrow \psi, I, w, \nu)))(\mathcal{V}(\psi \longrightarrow \varphi, I, w, \nu)) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$
iff $\mathcal{V}(\varphi \longrightarrow \psi, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$ and $\mathcal{V}(\psi \longrightarrow \varphi, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$
iff $(\mathcal{V}(\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{F}$ or $\mathcal{V}(\psi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T})$ and
$(\mathcal{V}(\psi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{F}$ or $\mathcal{V}(\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T})$,
for each $d_i \in \mathcal{D}_{\alpha_i} (i = 1, \ldots, n)$
iff $\mathcal{V}(\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathcal{V}(\psi, I, w, \nu)\ d_1 \ldots d_n$,
for each $d_i \in \mathcal{D}_{\alpha_i} (i = 1, \ldots, n)$
iff $\mathcal{V}(\varphi, I, w, \nu) = \mathcal{V}(\psi, I, w, \nu)$
iff $\mathcal{V}(\varphi = \psi, I, w, \nu) = \mathsf{T}$. [Proposition B.2.10]

□

Not surprisingly, the meaning of a rigid term is independent of the world, as shown by the next result.

Proposition B.2.12. *Let t be a term, I an interpretation, and ν a variable assignment. If t is rigid, then $\mathcal{V}(t, I, w, \nu) = \mathcal{V}(t, I, w', \nu)$, for all w, w'.*

Proof. The proof is by induction on the structure of t.

If t is a variable x, then $\mathcal{V}(x, I, w, \nu) = \nu(x) = \mathcal{V}(x, I, w', \nu)$, for all w, w'.

It t is a constant C, then C must be rigid, so that $\mathcal{V}(C, I, w, \nu) = V(C, w) = V(C, w') = \mathcal{V}(C, I, w, \nu)$, for all w, w'.

If t has the form $\lambda x.s$, then, for all w, w',

$\mathcal{V}(\lambda x.s, I, w, \nu)$
$=$ the function whose value for each $d \in \mathcal{D}_\alpha$ is $\mathcal{V}(s, I, w, \nu')$,
where x has type α and ν' is ν except $\nu'(x) = d$
$=$ the function whose value for each $d \in \mathcal{D}_\alpha$ is $\mathcal{V}(s, I, w', \nu')$ [Induction hypothesis]
$= \mathcal{V}(\lambda x.s, I, w', \nu)$.

If t has the form $(s\ r)$, then, for all w, w',

$\mathcal{V}((s\ r), I, w, \nu)$
$= \mathcal{V}(s, I, w, \nu)(\mathcal{V}(r, I, w, \nu))$
$= \mathcal{V}(s, I, w', \nu)(\mathcal{V}(r, I, w', \nu))$ [Induction hypothesis]
$= \mathcal{V}((s\ r), I, w', \nu)$.

If t has the form $(t_1, \ldots, t_n)$, the proof is similar to preceding case.
If t has the form $\Box_i s$, then, for all w, w',

$$\begin{aligned}
&\mathcal{V}(\Box_i s, I, w, \nu) \\
&= \mathcal{M}(\{\mathcal{V}(s, I, w_1, \nu)\}_{w_1 \in W_1}), \text{ where } W_1 = \{w_1 : w \, R_i \, w_1\} \\
&= \mathcal{M}(\{\mathcal{V}(s, I, w_2, \nu)\}_{w_2 \in W_2}), \text{ where } W_2 = \{w_2 : w' \, R_i \, w_2\} \quad \text{[Induction hypothesis]} \\
&= \mathcal{V}(\Box_i s, I, w', \nu).
\end{aligned}$$

□

Proposition B.2.13. *Let I be an interpretation, w a world in I, ν a variable assignment, and φ a formula. Then the following hold.*

1. $\mathcal{V}(\varphi = \top, I, w, \nu) = \mathcal{V}(\varphi, I, w, \nu)$.
2. $\mathcal{V}(\varphi = \bot, I, w, \nu) = \mathcal{V}(\neg\varphi, I, w, \nu)$.

Proof. 1.

$$\begin{aligned}
&\mathcal{V}(\varphi = \top, I, w, \nu) \\
&= \mathcal{V}((= \varphi), I, w, \nu) \; (\mathcal{V}(\top, I, w, \nu)) \\
&= (\mathcal{V}(=, I, w, \nu) \; (\mathcal{V}(\varphi, I, w, \nu))) \; (\mathcal{V}(\top, I, w, \nu)) \\
&= (V(=, w) \; (\mathcal{V}(\varphi, I, w, \nu))) \; (\mathcal{V}(\top, I, w, \nu)) \\
&= (V(=, w) \; (\mathcal{V}(\varphi, I, w, \nu))) \; (\mathsf{T}) \\
&= \mathcal{V}(\varphi, I, w, \nu).
\end{aligned}$$

2.

$$\begin{aligned}
&\mathcal{V}(\varphi = \bot, I, w, \nu) \\
&= \mathcal{V}((= \varphi), I, w, \nu) \; (\mathcal{V}(\bot, I, w, \nu)) \\
&= (\mathcal{V}(=, I, w, \nu) \; (\mathcal{V}(\varphi, I, w, \nu))) \; (\mathcal{V}(\bot, I, w, \nu)) \\
&= (V(=, w) \; (\mathcal{V}(\varphi, I, w, \nu))) \; (\mathcal{V}(\bot, I, w, \nu)) \\
&= (V(=, w) \; (\mathcal{V}(\varphi, I, w, \nu))) \; (\mathsf{F}) \\
&= \mathcal{V}(\neg\varphi, I, w, \nu).
\end{aligned}$$

□

The next result shows that the denotation of a *closed* term does not depend on the variable assignment.

Proposition B.2.14. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, V\rangle$ be an interpretation, $w \in W$, and t a term. If ν_1 and ν_2 are variable assignments with respect to $\{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}$ that agree on the free variables of t, then $\mathcal{V}(t, I, w, \nu_1) = \mathcal{V}(t, I, w, \nu_2)$.*

Proof. The proof is by induction on the structure of t.
If t is a variable x, then $\mathcal{V}(x, I, w, \nu_1) = \nu_1(x) = \nu_2(x) = \mathcal{V}(x, I, w, \nu_2)$.
If t is a constant C, then $\mathcal{V}(C, I, w, \nu_1) = V(C, w) = \mathcal{V}(C, I, w, \nu_2)$.

Let t be an abstraction $\lambda x.s$. Then

$$\begin{aligned}
&\mathcal{V}(\lambda x.s, I, w, \nu_1)\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(s, I, w, \nu_1'),\\
&\quad \text{where } x \text{ has type } \alpha \text{ and } \nu_1' \text{ is } \nu_1 \text{ except } \nu_1'(x) = d\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(s, I, w, \nu_2'),\\
&\quad \text{where } x \text{ has type } \alpha \text{ and } \nu_2' \text{ is } \nu_2 \text{ except } \nu_2'(x) = d\\
&\qquad\qquad \text{[Induction hypothesis, since } \nu_1' \text{ and } \nu_2' \text{ agree on the free variables of } s]\\
&= \mathcal{V}(\lambda x.s, I, w, \nu_2).
\end{aligned}$$

Let t be an application $(u\ v)$. Then

$$\begin{aligned}
&\mathcal{V}((u\ v), I, w, \nu_1)\\
&= \mathcal{V}(u, I, w, \nu_1)(\mathcal{V}(u, I, w, \nu_1))\\
&= \mathcal{V}(u, I, w, \nu_2)(\mathcal{V}(u, I, w, \nu_2)) \qquad \text{[Induction hypothesis]}\\
&= \mathcal{V}((u\ v), I, w, \nu_2).
\end{aligned}$$

Let t be a tuple $(t_1, \ldots, t_n)$. Then

$$\begin{aligned}
&\mathcal{V}((t_1, \ldots, t_n), I, w, \nu_1)\\
&= (\mathcal{V}(t_1, I, w, \nu_1), \ldots, \mathcal{V}(t_n, I, w, \nu_1))\\
&= (\mathcal{V}(t_1, I, w, \nu_2), \ldots, \mathcal{V}(t_n, I, w, \nu_2)) \qquad \text{[Induction hypothesis]}\\
&= \mathcal{V}((t_1, \ldots, t_n), I, w, \nu_2).
\end{aligned}$$

Let t have the form $\Box_i s$. Then

$$\begin{aligned}
&\mathcal{V}(\Box_i s, I, w, \nu_1)\\
&= \mathcal{M}(\{\mathcal{V}(s, I, w', \nu_1)\}_{w' \in W'})\\
&= \mathcal{M}(\{\mathcal{V}(s, I, w', \nu_2)\}_{w' \in W'}) \qquad \text{[Induction hypothesis]}\\
&= \mathcal{V}(\Box_i s, I, w, \nu_2).
\end{aligned}$$

□

Notation. If t is a closed term, then $\mathcal{V}(t, I, w)$ denotes $\mathcal{V}(t, I, w, \nu)$, for any variable assignment ν. Proposition B.2.14 shows that the value of $\mathcal{V}(t, I, w)$ does not depend on the choice of ν.

The next three results in this section about modalities will later provide useful assumptions for use in computations. The first concerns the case when the term t in $\Box_i t$ is rigid.

Proposition B.2.15. *Let t be a rigid term, I an interpretation, w a world in I, ν a variable assignment, and $i \in \{1, \ldots, m\}$. Then $\mathcal{V}(\Box_i t, I, w, \nu) = \mathcal{V}(t, I, w, \nu)$.*

Proof. Since t is rigid, by Proposition B.2.12, all the $\mathcal{V}(t, I, w', \nu)$, for $w' \in W'$, are equal. The result then follows directly from the definition of $\mathcal{V}(\Box_i t, I, w, \nu)$. □

The next proposition gives an important property of terms having type of the form $\alpha \to \beta$ that are applied to a rigid argument.

Proposition B.2.16. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, and $w \in W$. Let s be a term having type of the form $\alpha \to \beta$, t a term of type α that is rigid, and $i \in \{1, \ldots, m\}$. Then $\mathcal{V}((\Box_i s\ t), I, w, \nu) = \mathcal{V}(\Box_i(s\ t), I, w, \nu)$.*

Proof. The type of s has rank > 0. Thus

$$
\begin{aligned}
&\mathcal{V}((\Box_i s\ t), I, w, \nu)\\
&= \mathcal{V}(\Box_i s, I, w, \nu)(\mathcal{V}(t, I, w, \nu))\\
&= \mathcal{M}(\{\mathcal{V}(s, I, w', \nu)\}_{w'\in W'})(\mathcal{V}(t, I, w, \nu))\\
&= \mathcal{M}(\{\mathcal{V}(s, I, w', \nu)(\mathcal{V}(t, I, w, \nu))\}_{w'\in W'})\\
&= \mathcal{M}(\{\mathcal{V}(s, I, w', \nu)(\mathcal{V}(t, I, w', \nu))\}_{w'\in W'}) && [t \text{ is rigid}]\\
&= \mathcal{M}(\{\mathcal{V}((s\ t), I, w', \nu)\}_{w'\in W'})\\
&= \mathcal{V}(\Box_i(s\ t), I, w, \nu).
\end{aligned}
$$

□

Next it is shown that one can switch λ's and modalities.

Proposition B.2.17. *Let t be a term, $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, and $i \in \{1, \ldots, m\}$. Then $\mathcal{V}(\Box_i\lambda x.t, I, w, \nu) = \mathcal{V}(\lambda x.\Box_i t, I, w, \nu)$.*

Proof.

$$
\begin{aligned}
&\mathcal{V}(\Box_i\lambda x.t, I, w, \nu)\\
&= \mathcal{M}(\{\mathcal{V}(\lambda x.t, I, w', \nu)\}_{w'\in W'})\\
&= \mathcal{M}(\{\text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(t, I, w', \nu'),\\
&\qquad \text{where } x \text{ has type } \alpha \text{ and } \nu' \text{ is } \nu \text{ except } \nu'(x) = d\}_{w'\in W'})\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{M}(\{\mathcal{V}(t, I, w', \nu')\}_{w'\in W'}),\\
&\quad \text{where } x \text{ has type } \alpha \text{ and } \nu' \text{ is } \nu \text{ except } \nu'(x) = d\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(\Box_i t, I, w, \nu'),\\
&\quad \text{where } x \text{ has type } \alpha \text{ and } \nu' \text{ is } \nu \text{ except } \nu'(x) = d\\
&= \mathcal{V}(\lambda x.\Box_i t, I, w, \nu).
\end{aligned}
$$

□

There are analogous results to the preceding three for dual modalities.

Proposition B.2.18. *Let φ be a rigid biterm, I an interpretation, w a world in I, ν a variable assignment, and $i \in \{1, \ldots, m\}$. Then $\mathcal{V}(\Diamond_i\varphi, I, w, \nu) = \mathcal{V}(\varphi, I, w, \nu)$.*

Proof.

$$
\begin{aligned}
&\mathcal{V}(\Diamond_i\varphi, I, w, \nu)\\
&= \mathcal{V}(\neg\Box_i\neg\varphi, I, w, \nu)\\
&= V(\neg, w)(\mathcal{V}(\Box_i\neg\varphi, I, w, \nu))\\
&= V(\neg, w)(\mathcal{V}(\neg\varphi, I, w, \nu)) && \text{[Proposition B.2.15]}\\
&= \mathcal{V}(\neg\neg\varphi, I, w, \nu)\\
&= \mathcal{V}(\varphi, I, w, \nu) && \text{[Proposition B.2.1].}
\end{aligned}
$$

□

Proposition B.2.19. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, and $w \in W$. Let φ be a biterm having type of the form $\alpha \to \beta$, t a term of type α that is rigid, and $i \in \{1, \ldots, m\}$. Then $\mathcal{V}((\Diamond_i\varphi\ t), I, w, \nu) = \mathcal{V}(\Diamond_i(\varphi\ t), I, w, \nu)$.*

Proof.

$$
\begin{aligned}
&\mathcal{V}((\Diamond_i\varphi\ t), I, w, \nu)\\
&= \mathcal{V}((\neg\Box_i\neg\varphi\ t), I, w, \nu)\\
&= \mathcal{V}(\neg(\Box_i\neg\varphi\ t), I, w, \nu)) && \text{[Part 6 of Proposition B.2.1]}\\
&= V(\neg, w)(\mathcal{V}((\Box_i\neg\varphi\ t), I, w, \nu))\\
&= V(\neg, w)(\mathcal{V}(\Box_i(\neg\varphi\ t), I, w, \nu)) && \text{[Proposition B.2.16]}\\
&= V(\neg, w)(\mathcal{M}(\{\mathcal{V}((\neg\varphi\ t), I, w', \nu)\}_{w'\in W'}))\\
&= V(\neg, w)(\mathcal{M}(\{\mathcal{V}(\neg(\varphi\ t), I, w', \nu)\}_{w'\in W'}))\\
&= V(\neg, w)(\mathcal{V}(\Box_i\neg(\varphi\ t), I, w, \nu))\\
&= \mathcal{V}(\neg\Box_i\neg(\varphi\ t), I, w, \nu))\\
&= \mathcal{V}(\Diamond_i(\varphi\ t), I, w, \nu).
\end{aligned}
$$

□

Proposition B.2.20. *Let φ be a biterm, $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, and $i \in \{1, \ldots, m\}$. Then $\mathcal{V}(\Diamond_i\lambda x.\varphi, I, w, \nu) = \mathcal{V}(\lambda x.\Diamond_i\varphi, I, w, \nu)$.*

Proof.

$$
\begin{aligned}
&\mathcal{V}(\Diamond_i\lambda x.\varphi, I, w, \nu)\\
&= \mathcal{V}(\neg\Box_i\neg\lambda x.\varphi, I, w, \nu)\\
&= V(\neg, w)(\mathcal{V}(\Box_i\neg\lambda x.\varphi, I, w, \nu))\\
&= V(\neg, w)(\mathcal{M}(\{\mathcal{V}(\neg\lambda x.\varphi, I, w', \nu)\}_{w'\in W'}))\\
&= V(\neg, w)(\mathcal{M}(\{\mathcal{V}(\lambda x.\neg\varphi, I, w', \nu)\}_{w'\in W'})) && \text{[Part 9 of Proposition B.2.1]}\\
&= V(\neg, w)(\mathcal{V}(\Box_i\lambda x.\neg\varphi, I, w, \nu))\\
&= V(\neg, w)(\mathcal{V}(\lambda x.\Box_i\neg\varphi, I, w, \nu)) && \text{[Proposition B.2.17]}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathcal{V}(\neg\lambda x.\Box_i\neg\varphi, I, w, \nu) \\
&= \mathcal{V}(\lambda x.\neg\Box_i\neg\varphi, I, w, \nu) && \text{[Part 9 of Proposition B.2.1]} \\
&= \mathcal{V}(\lambda x.\Diamond_i\varphi, I, w, \nu).
\end{aligned}
$$

□

The next result establishes semantically the distribution axiom for the case of biterms.

Proposition B.2.21. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, and φ and ψ biterms of the same type. Then $\mathcal{V}(\Box_i(\varphi \longrightarrow \psi) \longrightarrow (\Box_i\varphi \longrightarrow \Box_i\psi), I, w, \nu) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}$.*

Proof. Suppose that the type of φ and ψ is $\alpha_1 \to \cdots \to \alpha_n \to o$. Let $d_j \in \mathcal{D}_{\alpha_j} (j = 1, \ldots, n)$. Then

$$
\begin{aligned}
&\mathcal{V}(\Box_i(\varphi \longrightarrow \psi) \longrightarrow (\Box_i\varphi \longrightarrow \Box_i\psi), I, w, \nu)\ d_1\ \ldots\ d_n \\
&= (V(\longrightarrow, w)(\mathcal{V}(\Box_i(\varphi \longrightarrow \psi), I, w, \nu)))(\mathcal{V}(\Box_i\varphi \longrightarrow \Box_i\psi, I, w, \nu))\ d_1\ \ldots\ d_n \\
&= \begin{cases} \mathsf{T} & \text{if } \mathcal{V}(\Box_i(\varphi \longrightarrow \psi), I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{F} \text{ or} \\ & \quad \mathcal{V}(\Box_i\varphi \longrightarrow \Box_i\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \\ \mathsf{F} & \text{otherwise} \end{cases} \\
&= \begin{cases} \mathsf{T} & \text{if } \mathcal{V}(\Box_i(\varphi \longrightarrow \psi), I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \text{ and } \mathcal{V}(\Box_i\varphi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \\ & \quad \text{implies } \mathcal{V}(\Box_i\psi, I, w, \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \\ \mathsf{F} & \text{otherwise} \end{cases} \\
&= \begin{cases} \mathsf{T} & \text{if } \mathcal{V}(\varphi \longrightarrow \psi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T} \text{ and } \mathcal{V}(\varphi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \\ & \quad \text{for each } w' \text{ such that } w\,R_i\,w', \\ & \quad \text{implies } \mathcal{V}(\psi, I, w', \nu)\ d_1\ \ldots\ d_n = \mathsf{T}, \text{ for each } w' \text{ such that } w\,R_i\,w' \\ \mathsf{F} & \text{otherwise} \end{cases} \\
&= \mathsf{T}.
\end{aligned}
$$

□

Proposition B.2.21 justifies the distribution axiom

$$\Box_i(\boldsymbol{\varphi} \longrightarrow \boldsymbol{\psi}) \longrightarrow (\Box_i\boldsymbol{\varphi} \longrightarrow \Box_i\boldsymbol{\psi}),$$

where $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are syntactical variables ranging over biterms of the same type, being used as a global assumption. (See below for the definition of a global assumption.)

Closely related to Proposition B.2.21 is the next result.

Proposition B.2.22. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, $w \in W$, and s and t terms of the same type. Then $\mathcal{V}(\Box_{j_1}\cdots\Box_{j_n}(s = t) \longrightarrow (\Box_{j_1}\cdots\Box_{j_n}s = \Box_{j_n}\cdots\Box_{j_n}t), I, w, \nu) = \mathsf{T}$.*

Proof. By Part 3 of Proposition B.2.1, it suffices to show that, if $\mathcal{V}(\Box_{j_1}\cdots\Box_{j_n}(s = t), I, w, \nu) = \mathsf{T}$, then $\mathcal{V}(\Box_{j_1}\cdots\Box_{j_n}s = \Box_{j_1}\cdots\Box_{j_n}t, I, w, \nu) = \mathsf{T}$. Suppose now that

$\mathcal{V}(\Box_{j_1} \cdots \Box_{j_n}(s = t), I, w, \nu) = \mathsf{T}$. Then, by Proposition B.2.7, $\mathcal{V}(s = t, I, w', \nu) = \mathsf{T}$, for each w' such that $w\, R_{j_1} \circ \cdots \circ R_{j_n}\, w'$. It follows that $\mathcal{V}(s, I, w', \nu) = \mathcal{V}(t, I, w', \nu)$, for each w' such that $w\, R_{j_1} \circ \cdots \circ R_{j_n}\, w'$. Hence $\mathcal{V}(\Box_{j_1} \cdots \Box_{j_n} s, I, w, \nu) = \mathcal{V}(\Box_{j_1} \cdots \Box_{j_n} t, I, w, \nu)$, and so $\mathcal{V}(\Box_{j_1} \cdots \Box_{j_n} s = \Box_{j_1} \cdots \Box_{j_n} t, I, w, \nu) = \mathsf{T}$. □

Proposition B.2.22 justifies the global assumption

$$\Box_{j_1} \cdots \Box_{j_n}(\boldsymbol{s} = \boldsymbol{t}) \longrightarrow (\Box_{j_1} \cdots \Box_{j_n} \boldsymbol{s} = \Box_{j_1} \cdots \Box_{j_n} \boldsymbol{t}),$$

where $\boldsymbol{s}$ and $\boldsymbol{t}$ are syntactical variables ranging over terms of the same type.

The next result gives a semantic version of η-reduction.

Proposition B.2.23. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}$, $w \in W$, t a term of type $\sigma \to \tau$, and x a variable of type σ that is not free in t. Then $\mathcal{V}(\lambda x.(t\ x) = t, I, w, \nu) = \mathsf{T}$.*

Proof.

$\mathcal{V}(\lambda x.(t\ x), I, w, \nu)$
$=$ the function whose value for each $d \in \mathcal{D}_\alpha$ is $\mathcal{V}((t\ x), I, w, \nu')$,
where ν' is ν except $\nu'(x) = d$
$=$ the function whose value for each $d \in \mathcal{D}_\alpha$ is $\mathcal{V}(t, I, w, \nu')(\mathcal{V}(x, I, w, \nu'))$
$=$ the function whose value for each $d \in \mathcal{D}_\alpha$ is $\mathcal{V}(t, I, w, \nu')(\nu'(x))$
$=$ the function whose value for each $d \in \mathcal{D}_\alpha$ is $\mathcal{V}(t, I, w, \nu')(d)$
$=$ the function whose value for each $d \in \mathcal{D}_\alpha$ is $\mathcal{V}(t, I, w, \nu)(d)$
[Proposition B.2.14, since x is not free in t so that ν and ν' agree on the free variables of t]
$= \mathcal{V}(t, I, w, \nu)$.

Hence $\mathcal{V}(\lambda x.(t\ x) = t, I, w, \nu) = \mathsf{T}$. □

Proposition B.2.23 justifies the global assumption

$$\forall f.(\lambda x.(f\ x) = f).$$

This proposition also provides justification for the rule of η-reduction.

Definition B.2.12. The rule of *η-reduction* is as follows: $\lambda x.(t\ x) \succ_\eta t$, if x is not free in t.

Function definitions are used by agents in their belief bases. In the logic, these definitions usually have the form either $\Box_{j_1} \cdots \Box_{j_n} \forall x.((f\ x) = t)$ or $\Box_{j_1} \cdots \Box_{j_n}(f = \lambda x.t)$. The next result shows that semantically these two formulas are the same and hence can be used interchangeably.

Proposition B.2.24. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}$, $w \in W$, f a constant having signature $\sigma \to \tau$, and t a term of type τ. Then*

$$\mathcal{V}(\Box_{j_1} \cdots \Box_{j_n} \forall x.((f\ x) = t), I, w, \nu) = \mathcal{V}(\Box_{j_1} \cdots \Box_{j_n}(f = \lambda x.t), I, w, \nu).$$

Proof. Let $w' \in W$. Then

$$\begin{aligned}
&\mathcal{V}(\forall x.((f\ x) = t), I, w', \nu) = \mathsf{T} \\
&\text{iff } \mathcal{V}((f\ x) = t, I, w', \nu') = \mathsf{T}, \text{for every } x\text{-variant } \nu' \text{ of } \nu \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[Part 3 of Proposition B.2.4]} \\
&\text{iff } \mathcal{V}((f\ x), I, w', \nu') = \mathcal{V}(t, I, w', \nu'), \text{for every } x\text{-variant } \nu' \text{ of } \nu \quad \text{[Proposition B.2.10]} \\
&\text{iff } \mathcal{V}(\lambda x.(f\ x), I, w', \nu) = \mathcal{V}(\lambda x.t, I, w', \nu) \\
&\text{iff } \mathcal{V}(f, I, w', \nu) = \mathcal{V}(\lambda x.t, I, w', \nu) \qquad\qquad\qquad\qquad\qquad\qquad \text{[Proposition B.2.23]} \\
&\text{iff } \mathcal{V}(f = \lambda x.t, I, w', \nu) = \mathsf{T}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[Proposition B.2.10]}
\end{aligned}$$

It follows from Proposition B.2.5 that

$$\mathcal{V}(\Box_{j_1} \cdots \Box_{j_n} \forall x.((f\ x) = t), I, w, \nu) = \mathsf{T} \text{ iff } \mathcal{V}(\Box_{j_1} \cdots \Box_{j_n} (f = \lambda x.t), I, w, \nu) = \mathsf{T}.$$

Hence $\mathcal{V}(\Box_{j_1} \cdots \Box_{j_n} \forall x.((f\ x) = t), I, w, \nu) = \mathcal{V}(\Box_{j_1} \cdots \Box_{j_n} (f = \lambda x.t), I, w, \nu)$. □

The next result is concerned with equality of functions.

Proposition B.2.25. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}$, $w \in W$, s and t terms having type of the form $\sigma \to \tau$, and x a variable of type σ not occurring freely in s or t. Then $\mathcal{V}((s = t) = \forall x.((s\ x) = (t\ x)), I, w, \nu) = \mathsf{T}$.*

Proof.

$$\begin{aligned}
&\mathcal{V}(\forall x.((s\ x) = (t\ x)), I, w, \nu) = \mathsf{T} \\
&\text{iff } \mathcal{V}((s\ x) = (t\ x), I, w, \nu') = \mathsf{T}, \text{ for every } x\text{-variant } \nu' \text{ of } \nu \\
&\text{iff } \mathcal{V}((s\ x), I, w, \nu') = \mathcal{V}((t\ x), I, w, \nu'), \text{ for every } x\text{-variant } \nu' \text{ of } \nu \\
&\text{iff } \mathcal{V}(s, I, w, \nu)(\nu'(x)) = \mathcal{V}(t, I, w, \nu)(\nu'(x)), \text{ for every } x\text{-variant } \nu' \text{ of } \nu \\
&\text{iff } \mathcal{V}(s, I, w, \nu) = \mathcal{V}(t, I, w, \nu) \\
&\text{iff } \mathcal{V}(s = t, I, w, \nu) = \mathsf{T}.
\end{aligned}$$

Hence $\mathcal{V}((s = t) = \forall x.((s\ x) = (t\ x)), I, w, \nu) = \mathsf{T}$. □

Proposition B.2.25 justifies the axiom of extensionality

$$\forall f.\forall g.((f = g) = \forall x.((f\ x) = (g\ x)))$$

being used as a global assumption.

The next result justifies a global assumption concerning tuples.

Proposition B.2.26. *Let $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, V\rangle$ be an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}$, $w \in W$, and $(s_1, \ldots, s_n)$ and $(t_1, \ldots, t_n)$ terms of the same type. Then $\mathcal{V}(((s_1, \ldots, s_n) = (t_1, \ldots, t_n)) = (s_1 = t_1) \wedge \cdots \wedge (s_n = t_n), I, w, \nu) = \mathsf{T}$.*

Proof.

$$\begin{aligned}
&\mathcal{V}((s_1,\ldots,s_n)=(t_1,\ldots,t_n),I,w,\nu)=\mathsf{T}\\
&\text{iff } \mathcal{V}((s_1,\ldots,s_n),I,w,\nu)=\mathcal{V}((t_1,\ldots,t_n),I,w,\nu)\\
&\text{iff } (\mathcal{V}(s_1,I,w,\nu),\ldots,\mathcal{V}(s_n,I,w,\nu))=(\mathcal{V}(t_1,I,w,\nu),\ldots,\mathcal{V}(t_n,I,w,\nu))\\
&\text{iff } \mathcal{V}(s_i,I,w,\nu)=\mathcal{V}(t_i,I,w,\nu),\ \text{for } i=1,\ldots,n\\
&\text{iff } \mathcal{V}(s_i=t_i,I,w,\nu)=\mathsf{T},\ \text{for } i=1,\ldots,n\\
&\text{iff } \mathcal{V}((s_1=t_1)\wedge\cdots\wedge(s_n=t_n),I,w,\nu)=\mathsf{T}.
\end{aligned}$$

Hence $\mathcal{V}(((s_1,\ldots,s_n)=(t_1,\ldots,t_n))=(s_1=t_1)\wedge\cdots\wedge(s_n=t_n),I,w,\nu)=\mathsf{T}$. □

Proposition B.2.26 justifies

$$\forall x_1.\ \ldots\ \forall x_n.\forall y_1.\ \ldots\ \forall y_n.(((x_1,\ldots,x_n)=(y_1,\ldots,y_n))=(x_1=y_1)\wedge\cdots\wedge(x_n=y_n))$$

being used as a global assumption.

B.2.3 Admissible Substitutions

The next few results are concerned with knowing the semantics of a term after a substitution is applied or a subterm replaced. For these, a restriction on the application of substitutions is needed.

Definition B.2.13. Let t be a term and $\theta \triangleq \{x_1/t_1,\ldots,x_n/t_n\}$ a substitution. Then θ is *admissible* with respect to t if, for $i=1,\ldots,n$, whenever x_i has a free occurrence in t that is a modal occurrence, t_i is rigid.

Proposition B.2.27. *Let t be a term, $\theta \triangleq \{x_1/t_1,\ldots,x_n/t_n\}$ a substitution, I an interpretation, w a world in I, and ν a variable assignment. Suppose that θ is admissible with respect to t. Then $\mathcal{V}(t\theta,I,w,\nu)=\mathcal{V}(t,I,w,\nu')$, where $\nu'(x_i)=\mathcal{V}(t_i,I,w,\nu)$, for $i=1,\ldots,n$, and $\nu'(y)=\nu(y)$, for $y\notin\{x_1,\ldots,x_n\}$.*

Proof. It can be supposed without loss of generality that x_i is free in t, for $i=1,\ldots,n$. For, suppose $x_{i_1},\ldots,x_{i_k}$ are the variables amongst the x_i that are free in t, so that $\theta|_t=\{x_{i_1}/t_{i_1},\ldots,x_{i_k}/t_{i_k}\}$. Then

$$\begin{aligned}
&\mathcal{V}(t\theta,I,w,\nu)\\
&=\mathcal{V}(t\theta|_t,I,w,\nu) && \text{[Proposition B.1.6]}\\
&=\mathcal{V}(t,I,w,\nu^*),\ \text{where } \nu^*(x_{i_j})=\mathcal{V}(t_{i_j},I,w,\nu),\ \text{for } j=1,\ldots,k,\\
&\quad\text{and } \nu^*(y)=\nu(y),\ \text{for } y\notin\{x_{i_1},\ldots,x_{i_k}\} && \text{[Assumption]}\\
&=\mathcal{V}(t,I,w,\nu'),\ \text{where } \nu'(x_i)=\mathcal{V}(t_i,I,w,\nu),\ \text{for } i=1,\ldots,n,\\
&\quad\text{and } \nu'(y)=\nu(y),\ \text{for } y\notin\{x_1,\ldots,x_n\}.\\
&\qquad\qquad\text{[Proposition B.2.14, since } \nu^* \text{ and } \nu' \text{ agree on the free variables in } t]
\end{aligned}$$

The proof is by induction on the structure of t.

Suppose t is a variable x. If x is distinct from all the x_i, then $\mathcal{V}(x\theta,I,w,\nu)=\mathcal{V}(x,I,w,\nu)=\mathcal{V}(x,I,w,\nu')$, since ν and ν' agree on the free variables of x. If x is

x_i, for some $i \in \{1, \ldots, n\}$, then $\mathcal{V}(x_i\theta, I, w, \nu) = \mathcal{V}(t_i, I, w, \nu) = \mathcal{V}(x_i, I, w, \nu')$, since $\nu'(x_i) = \mathcal{V}(t_i, I, w, \nu)$.

Suppose t is a constant C. Then $\mathcal{V}(C\theta, I, w, \nu) = \mathcal{V}(C, I, w, \nu) = \mathcal{V}(C, I, w, \nu')$.

Suppose t is an abstraction. Without loss of generality, it can be assumed that, for $i = 1, \ldots, n$, x_i is free in t.

(a) Suppose that t is an abstraction $\lambda x.s$ of type $\alpha \to \beta$ such that, for all $i \in \{1, \ldots, n\}$, x is not free in t_i. Then

$$
\begin{aligned}
&\mathcal{V}((\lambda x.s)\theta, I, w, \nu) \\
&= \mathcal{V}(\lambda x.(s\theta), I, w, \nu) && \text{[Proposition B.1.8]} \\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(s\theta, I, w, \overline{\nu}), \\
&\quad \text{where } \overline{\nu} \text{ is } \nu \text{ except } \overline{\nu}(x) = d \\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(s, I, w, \overline{\nu'}), \text{ where} \\
&\quad \overline{\nu'}(x_i) = \mathcal{V}(t_i, I, w, \overline{\nu}), \text{ for all } i, \text{ and } \overline{\nu'}(y) = \overline{\nu}(y), \text{ for } y \notin \{x_1, \ldots, x_n\} \\
&\qquad\qquad \text{[Induction hypothesis]} \\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(s, I, w, \overline{\nu'}), \\
&\quad \text{where } \overline{\nu'} \text{ is } \nu' \text{ except } \overline{\nu'}(x) = d, \text{ and } \nu'(x_i) = \mathcal{V}(t_i, I, w, \nu), \\
&\quad \text{for } i = 1, \ldots, n, \text{ and } \nu'(y) = \nu(y), \text{ for } y \notin \{x_1, \ldots, x_n\} \\
&\qquad \text{[Since } x \text{ is not free in any } t_i, \ \nu \text{ and } \overline{\nu} \text{ agree on the free variables} \\
&\qquad\ \text{in each } t_i; \text{ hence } \mathcal{V}(t_i, I, w, \nu) = \mathcal{V}(t_i, I, w, \overline{\nu}); \text{ thus } \overline{\nu'} = \overline{\nu'}\ ] \\
&= \mathcal{V}(\lambda x.s, I, w, \nu').
\end{aligned}
$$

(b) Suppose that t is an abstraction $\lambda x.s$ of type $\alpha \to \beta$ such that, for some $i \in \{1, \ldots, n\}$, x is free in t_i. Then

$$
\begin{aligned}
&\mathcal{V}((\lambda x.s)\theta, I, w, \nu) \\
&= \mathcal{V}(\lambda y.(s(\{x/y\} \cup \theta)), I, w, \nu) \quad \text{[Proposition B.1.8]} \\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(s(\{x/y\} \cup \theta), I, w, \overline{\nu}), \\
&\quad \text{where } \overline{\nu} \text{ is } \nu \text{ except } \overline{\nu}(y) = d \\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(s, I, w, \overline{\nu'}), \text{ where} \\
&\quad \overline{\nu'}(x) = \mathcal{V}(y, I, w, \overline{\nu}), \ \overline{\nu'}(x_i) = \mathcal{V}(t_i, I, w, \overline{\nu}), \text{ for } i = 1, \ldots, n, \text{ and} \\
&\quad \overline{\nu'}(z) = \overline{\nu}(z), \text{ for } z \notin (\{x_1, \ldots, x_n\} \cup \{x\}) \qquad \text{[Induction hypothesis]} \\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(s, I, w, \overline{\nu'}), \\
&\quad \text{where } \overline{\nu'} \text{ is } \nu' \text{ except } \overline{\nu'}(x) = d, \text{ and } \nu'(x_i) = \mathcal{V}(t_i, I, w, \nu), \\
&\quad \text{for } i = 1, \ldots, n, \text{ and } \nu'(z) = \nu(z), \text{ for } z \notin \{x_1, \ldots, x_n\} \\
&\qquad\qquad\qquad\qquad [\overline{\nu'} \text{ and } \overline{\nu'} \text{ agree, except possibly on } y] \\
&= \mathcal{V}(\lambda x.s, I, w, \nu').
\end{aligned}
$$

Suppose t has the form $(u\ v)$. Then

$$
\begin{aligned}
&\mathcal{V}((u\ v)\theta, I, w, \nu) \\
&= \mathcal{V}((u\theta\ v\theta), I, w, \nu)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathcal{V}(u\theta, I, w, \nu)(\mathcal{V}(v\theta, I, w, \nu)) \\
&= \mathcal{V}(u, I, w, \nu')(\mathcal{V}(v, I, w, \nu')) && \text{[Induction hypothesis]} \\
&= \mathcal{V}((u\ v), I, w, \nu'),
\end{aligned}
$$

where $\nu'(x_i) = \mathcal{V}(t_i, I, w, \nu)$, for $i = 1, \ldots, n$, and $\nu'(y) = \nu(y)$, for $y \notin \{x_1, \ldots, x_n\}$.

Suppose that t has the form $(t_1, \ldots, t_n)$. Then

$$
\begin{aligned}
&\mathcal{V}((t_1, \ldots, t_n)\theta, I, w, \nu) \\
&= \mathcal{V}((t_1\theta, \ldots, t_n\theta), I, w, \nu) \\
&= (\mathcal{V}(t_1\theta, I, w, \nu), \ldots, (\mathcal{V}(t_n\theta, I, w, \nu)) \\
&= (\mathcal{V}(t_1, I, w, \nu'), \ldots, (\mathcal{V}(t_n, I, w, \nu')) && \text{[Induction hypothesis]} \\
&= \mathcal{V}((t_1, \ldots, t_n), I, w, \nu'),
\end{aligned}
$$

where $\nu'(x_i) = \mathcal{V}(t_i, I, w, \nu)$, for $i = 1, \ldots, n$, and $\nu'(y) = \nu(y)$, for $y \notin \{x_1, \ldots, x_n\}$.

Suppose that t has the form $\Box_i s$. Then

$$
\begin{aligned}
&\mathcal{V}((\Box_i s)\theta, I, w, \nu) \\
&= \mathcal{V}(\Box_i(s\theta), I, w, \nu) \\
&= \mathcal{M}(\{\mathcal{V}(s\theta, I, w', \nu)\}_{w' \in W'}) \\
&= \mathcal{M}(\{\mathcal{V}(s, I, w', \nu') : \nu'(x_i) = \mathcal{V}(t_i, I, w', \nu), \text{ for } i = 1, \ldots, n, \text{ and} \\
&\quad \nu'(y) = \nu(y), \text{ for } y \notin \{x_1, \ldots, x_n\}\}_{w' \in W'}) && \text{[Induction hypothesis]} \\
&= \mathcal{M}(\{\mathcal{V}(s, I, w', \nu')\}_{w' \in W'}), \\
&\quad \text{where } \nu'(x_i) = \mathcal{V}(t_i, I, w, \nu), \text{ for } i = 1, \ldots, n, \text{ and } \nu'(y) = \nu(y), \\
&\quad \text{for } y \notin \{x_1, \ldots, x_n\} && [\theta \text{ is admissible w.r.t. } t] \\
&= \mathcal{V}(\Box_i s, I, w, \nu'), \\
&\quad \text{where } \nu'(x_i) = \mathcal{V}(t_i, I, w, \nu), \text{ for } i = 1, \ldots, n, \text{ and } \nu'(y) = \nu(y), \\
&\quad \text{for } y \notin \{x_1, \ldots, x_n\}.
\end{aligned}
$$

□

The next example shows that the condition that θ be admissible with respect to t in Proposition B.2.27 cannot be dropped. The reason for this comes from a basic asymmetry in the treatment of constants and variables – the meaning of constants can change from world to world, while each variable assignment assigns the same meaning to a variable in each world.

Example B.2.1. Let α be a type, $p : \alpha \to o$, $C : \alpha$, $t \triangleq \Box_i(p\ x)$, and $\theta \triangleq \{x/C\}$. Also let I be an interpretation with two worlds w and w', an accessibility relation R_i such that $w\ R_i\ w'$, domain $\mathfrak{D}_\alpha \triangleq \{a, b\}$, and mapping V, where $V(p, w) = V(p, w') =$ relation that is true on a and false on b, and $V(C, w) = a$ and $V(C, w') = b$. Finally, let ν be any variable assignment that maps x to a. Since C is not rigid, θ is not admissible with respect to t.

Then $\mathcal{V}(t\theta, I, w, \nu) = \mathcal{V}(\Box_i(p\ C), I, w, \nu) = \mathsf{F}$. But $\mathcal{V}(t, I, w, \nu') = \mathcal{V}(\Box_i(p\ x), I, w, \nu') = \mathsf{T}$, where $\nu'(x) = \mathcal{V}(C, I, w, \nu) = V(C, w) = a$ and $\nu'(y) = \nu(y)$, for $y \neq x$.

Since variables are rigid, the next result is a special case of Proposition B.2.27.

Proposition B.2.28. *Let t be a term, I an interpretation, w a world in I, ν a variable assignment, and x and y variables. Then $\mathcal{V}(t\{x/y\}, I, w, \nu) = \mathcal{V}(t, I, w, \nu')$, where $\nu'(x) = \nu(y)$, and $\nu'(z) = \nu(z)$, for $z \neq x$.*

Proposition B.2.29. *Let t be a term, s and r terms of the same type, I an interpretation, w a world in I, and ν a variable assignment. Suppose that $\{x/s\}$ and $\{x/r\}$ are admissible with respect to t. If $\mathcal{V}(s = r, I, w, \nu) = \mathsf{T}$, then $\mathcal{V}(t\{x/s\}, I, w, \nu) = \mathcal{V}(t\{x/r\}, I, w, \nu)$.*

Proof. Proposition B.2.10 shows that $\mathcal{V}(s, I, w, \nu) = \mathcal{V}(r, I, w, \nu)$. Then

$$\begin{aligned}
&\mathcal{V}(t\{x/s\}, I, w, \nu) \\
&= \mathcal{V}(t, I, w, \nu'), \text{ where } \nu' \text{ is } \nu \text{ except that } \nu'(x) = \mathcal{V}(s, I, w, \nu) && \text{[Proposition B.2.27]} \\
&= \mathcal{V}(t, I, w, \nu'), \text{ where } \nu' \text{ is } \nu \text{ except that } \nu'(x) = \mathcal{V}(r, I, w, \nu) \\
&= \mathcal{V}(t\{x/r\}, I, w, \nu). && \text{[Proposition B.2.27]}
\end{aligned}$$

□

Proposition B.2.30. *Let φ be a biterm, I an interpretation, w a world in I, and θ a substitution. Suppose that θ is admissible with respect to φ and that $\mathcal{V}(\Box_{k_1} \ldots \Box_{k_m} \varphi, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$, for each variable assignment ν. Then $\mathcal{V}(\Box_{k_1} \ldots \Box_{k_m} \varphi\theta, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$, for each variable assignment ν.*

Proof. The proof is by induction on the length m of $\Box_{k_1} \ldots \Box_{k_m}$.

Suppose first that $m = 0$. Assume that $\mathcal{V}(\varphi, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$, for each variable assignment ν. Let θ be $\{x_1/t_1, \ldots, x_n/t_n\}$ and ν a variable assignment. By Proposition B.2.27, $\mathcal{V}(\varphi\theta, I, w, \nu) = \mathcal{V}(\varphi, I, w, \nu')$, where $\nu'(x_i) = \mathcal{V}(t_i, I, w, \nu)$, for $i = 1, \ldots, n$, and $\nu'(y) = \nu(y)$, for $y \notin \{x_1, \ldots, x_n\}$. Hence $\mathcal{V}(\varphi\theta, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$.

Suppose next that the result holds for length m. Assume that $\mathcal{V}(\Box_{k_1} \ldots \Box_{k_{m+1}} \varphi, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$, for each variable assignment ν. Then it follows that, for each variable assignment ν, $\mathcal{V}(\Box_{k_2} \ldots \Box_{k_{m+1}} \varphi, I, w', \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$, for each w' such that $w\, R_{k_1}\, w'$. By the induction hypothesis, for each variable assignment ν, $\mathcal{V}(\Box_{k_2} \ldots \Box_{k_{m+1}} \varphi\theta, I, w', \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$, for each w' such that $w\, R_{k_1}\, w'$. Thus, for each variable assignment ν, $\mathcal{V}(\Box_{k_1} \ldots \Box_{k_{m+1}} \varphi\theta, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$. □

B.2.4 Term Replacement and Denotations

The next result will be important for the computational mechanism introduced in Appendix B.3.

Proposition B.2.31. *Let t be a term, s a subterm of t at occurrence o, r a term having the same type as s, I an interpretation, and w a world in I. Suppose that $k_1 \ldots k_m$ is the modal path to o in t and $\mathcal{V}(\Box_{k_1} \cdots \Box_{k_m}(s = r), I, w, \nu) = \mathsf{T}$, for each variable assignment ν. Then $\mathcal{V}(t = t[s/r]_o, I, w, \nu) = \mathsf{T}$, for each variable assignment ν.*

Proof. The proof is by induction on the length n of o.

Suppose first that $n = 0$. Thus t is s, $t[s/r]_o$ is r, and the modal path to o is empty. Suppose that ν is a variable assignment. Then $\mathcal{V}(t = t[s/r]_o, I, w, \nu) = \mathcal{V}(s = r, I, w, \nu) = \mathsf{T}$.

Suppose next that the result holds for occurrences of length n and o has length $n+1$. Thus t has the form $\lambda x.q$, $(u\ v)$, $(t_1, \ldots, t_n)$, or $\Box_i q$.

Consider the case when t has the form $\lambda x.q$ and $o = 1o'$, for some o'. Suppose that ν is a variable assignment. Then

$$\begin{aligned}
&\mathcal{V}(\lambda x.q, I, w, \nu)\\
&= \text{the function whose value for each } d \in \mathcal{D}_\gamma \text{ is } \mathcal{V}(q, I, w, \nu'),\\
&\quad \text{where the type of } t \text{ is } \gamma \to \delta \text{ and } \nu' \text{ is } \nu \text{ except } \nu'(x) = d\\
&= \text{the function whose value for each } d \in \mathcal{D}_\gamma \text{ is } \mathcal{V}(q[s/r]_{o'}, I, w, \nu')\\
&\qquad\qquad \text{[Induction hypothesis (since modal path to } o' \text{ is same as to } o) \text{ and}\\
&\qquad\qquad \text{Proposition B.2.10]}\\
&= \mathcal{V}(\lambda x.(q[s/r]_{o'}), I, w, \nu)\\
&= \mathcal{V}((\lambda x.q)[s/r]_o, I, w, \nu).
\end{aligned}$$

By Proposition B.2.10, $\mathcal{V}(\lambda x.q = (\lambda x.q)[s/r]_o, I, w, \nu) = \mathsf{T}$.

Consider next the case when t has the form $(u\ v)$ and $o = 1o'$, for some o'. Suppose that ν is a variable assignment. Then

$$\begin{aligned}
&\mathcal{V}((u\ v), I, w, \nu)\\
&= \mathcal{V}(u, I, w, \nu)(\mathcal{V}(v, I, w, \nu))\\
&= \mathcal{V}(u[s/r]_{o'}, I, w, \nu)(\mathcal{V}(v, I, w, \nu)) \quad \text{[Induction hypothesis and Proposition B.2.10]}\\
&= \mathcal{V}((u[s/r]_{o'}\ v), I, w, \nu)\\
&= \mathcal{V}((u\ v)[s/r]_o, I, w, \nu).
\end{aligned}$$

By Proposition B.2.10, $\mathcal{V}((u\ v) = (u\ v)[s/r]_o, I, w, \nu) = \mathsf{T}$. The case when $o = 2o'$ is similar.

The case when t has the form $(t_1, \ldots, t_n)$ is similar to the preceding one.

Consider finally the case when t has the form $\Box_i q$. Thus $i = k_1$, $o = 1o'$, for some o', and the modal path to o' in q is $k_2 \ldots k_m$. Furthermore, for each variable assignment ν and each w' such that $w\, R_i\, w'$, $\mathcal{V}(\Box_{k_2} \cdots \Box_{k_m}(s = r), I, w', \nu) = \mathsf{T}$. Suppose that ν is a variable assignment. Then

$$\begin{aligned}
&\mathcal{V}(\Box_i q, I, w, \nu)\\
&= \mathcal{M}(\{\mathcal{V}(q, I, w', \nu)\}_{w' \in W'})\\
&= \mathcal{M}(\{\mathcal{V}(q[s/r]_{o'}, I, w', \nu)\}_{w' \in W'}) \quad \text{[Induction hypothesis and Proposition B.2.10]}\\
&= \mathcal{V}(\Box_i(q[s/r]_{o'}, I, w, \nu)\\
&= \mathcal{V}((\Box_i q)[s/r]_o, I, w, \nu).
\end{aligned}$$

By Proposition B.2.10, $\mathcal{V}(\Box_i q = (\Box_i q)[s/r]_o, I, w, \nu) = \mathsf{T}$. □

B.2.5 Denotations of α-equivalent Terms

The next result establishes that α-equivalent terms have the same meaning.

Proposition B.2.32. *Let s and t be terms, $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, and $w \in W$. Then the following hold.*

1. *If y is not free in t, then $\mathcal{V}(\lambda x.t, I, w, \nu) = \mathcal{V}(\lambda y.(t\{x/y\}), I, w, \nu)$.*
2. *If $s \longrightarrow_\alpha t$, then $\mathcal{V}(s, I, w, \nu) = \mathcal{V}(t, I, w, \nu)$.*
3. *If $s \overset{*}{\longleftrightarrow}_\alpha t$, then $\mathcal{V}(s, I, w, \nu) = \mathcal{V}(t, I, w, \nu)$.*

Proof. 1. Suppose that x and y are variables of type α. Then

$$
\begin{aligned}
&\mathcal{V}(\lambda y.(t\{x/y\}), I, w, \nu)\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(t\{x/y\}, I, w, \nu'),\\
&\quad \text{where } \nu' \text{ is } \nu \text{ except } \nu'(y) = d\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(t, I, w, \nu''), \text{ where}\\
&\quad \nu''(x) = \nu'(y),\ \nu''(z) = \nu'(z), \text{ for } z \neq x, \text{ and } \nu' \text{ is } \nu \text{ except } \nu'(y) = d\\
&\qquad\qquad \text{[Proposition B.2.28]}\\
&= \text{the function whose value for each } d \in \mathcal{D}_\alpha \text{ is } \mathcal{V}(t, I, w, \nu'''),\\
&\quad \text{where } \nu''' \text{ is } \nu \text{ except } \nu'''(x) = d\\
&\qquad\qquad \text{[Proposition B.2.14, since } y \text{ is not free in } t \text{ so } \nu'' \text{ and } \nu''' \text{ agree on}\\
&\qquad\qquad \text{the free variables of } t\text{]}\\
&= \mathcal{V}(\lambda x.t, I, w, \nu).
\end{aligned}
$$

2. Suppose that t is $s[u/v]_o$, where $u \succ_\alpha v$. From Part 1, $\mathcal{V}(u, I, w, \nu) = \mathcal{V}(v, I, w, \nu)$, for each world w and variable assignment ν. By Proposition B.2.10, $\mathcal{V}(u = v, I, w, \nu) = \mathsf{T}$, for each world w and variable assignment ν. Now let $k_1 \ldots k_m$ be the modal path to o in s. Thus $\mathcal{V}(\Box_{k_1} \cdots \Box_{k_m}(u = v), I, w, \nu) = \mathsf{T}$, for each world w and variable assignment ν. By Proposition B.2.31, $\mathcal{V}(s = s[u/v]_o, I, w, \nu) = \mathsf{T}$, for each world w and variable assignment ν. Hence, by Proposition B.2.10 again, $\mathcal{V}(s, I, w, \nu) = \mathcal{V}(s[u/v]_o, I, w, \nu)$, for each world w and variable assignment ν.

3. This is an easy induction on the length of the path of α-conversions from s to t, using Part 2. □

As shown by the next example, the condition that y not be free in t in Part 1 of Proposition B.2.32 cannot be dropped.

Example B.2.2. Let x and y be variables of type *Nat*, I an interpretation consisting of a single world w whose domain for the type *Nat* is the set of natural numbers $\mathbb{N}_0$, and ν a variable assignment that maps y to 0. Then

$$
\begin{aligned}
&\mathcal{V}(\lambda x.y, I, w, \nu)\\
&= \text{the function from } \mathbb{N}_0 \text{ to } \mathbb{N}_0 \text{ that maps each natural number to 0.}
\end{aligned}
$$

But

$$
\begin{aligned}
&\mathcal{V}(\lambda y.(y\{x/y\}), I, w, \nu)\\
&= \mathcal{V}(\lambda y.y, I, w, \nu)\\
&= \text{the identity function on } \mathbb{N}_0.
\end{aligned}
$$

B.2.6 β-Reduction

The relation $\succ_\beta$ corresponding to β-reduction is now defined.

Definition B.2.14. The rule of *β-reduction* is as follows: $(\lambda x.s\ t)\ \succ_\beta\ s\{x/t\}$, if $\{x/t\}$ is admissible with respect to s.

Definition B.2.15. The relation $\longrightarrow_\beta$ is defined by $u \longrightarrow_\beta u[s/t]_o$ if s is a subterm of u at occurrence o and $s\ \succ_\beta\ t$.

Let $\xrightarrow{*}_\beta$ be the reflexive, transitive closure of $\longrightarrow_\beta$.

Let $\xleftrightarrow{*}_\beta$ be the reflexive, symmetric, and transitive closure of $\longrightarrow_\beta$.

Definition B.2.16. If $s \xleftrightarrow{*}_\beta t$, then s and t are said to be *β-equivalent*.

The next result establishes that β-equivalent terms have the same meaning.

Proposition B.2.33. *Let s and t be terms, $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ an interpretation, ν a variable assignment with respect to $\{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}$, and $w \in W$. Then the following hold.*

1. *If $\{x/t\}$ is admissible with respect to s, then $\mathcal{V}((\lambda x.s\ t), I, w, \nu) = \mathcal{V}(s\{x/t\}, I, w, \nu)$.*
2. *If $s \longrightarrow_\beta t$, then $\mathcal{V}(s, I, w, \nu) = \mathcal{V}(t, I, w, \nu)$.*
3. *If $s \xleftrightarrow{*}_\beta t$, then $\mathcal{V}(s, I, w, \nu) = \mathcal{V}(t, I, w, \nu)$.*

Proof. 1.

$$\begin{aligned}
&\mathcal{V}((\lambda x.s\ t), I, w, \nu)\\
&= \mathcal{V}(\lambda x.s, I, w, \nu)\ (\mathcal{V}(t, I, w, \nu))\\
&= \mathcal{V}(s, I, w, \nu'), \text{where } \nu' \text{ is } \nu \text{ except that } \nu'(x) = \mathcal{V}(t, I, w, \nu)\\
&= \mathcal{V}(s\{x/t\}, I, w, \nu). \qquad \text{[Proposition B.2.27]}
\end{aligned}$$

2. Suppose that t is $s[u/v]_o$, where $u \succ_\beta v$. From Part 1, $\mathcal{V}(u, I, w, \nu) = \mathcal{V}(v, I, w, \nu)$, for each world w and variable assignment ν. By Proposition B.2.10, $\mathcal{V}(u = v, I, w, \nu) = \mathsf{T}$, for each world w and variable assignment ν. Now let $k_1 \ldots k_m$ be the modal path to o in s. Thus $\mathcal{V}(\Box_{k_1} \cdots \Box_{k_m}(u = v), I, w, \nu) = \mathsf{T}$, for each world w and variable assignment ν. By Proposition B.2.31, $\mathcal{V}(s = s[u/v]_o, I, w, \nu) = \mathsf{T}$, for each world w and variable assignment ν. Hence, by Proposition B.2.10 again, $\mathcal{V}(s, I, w, \nu) = \mathcal{V}(s[u/v]_o, I, w, \nu)$, for each world w and variable assignment ν.

3. This is an easy induction on the length of the path of β-reductions from s to t, using Part 2. □

The next example shows that the condition that $\{x/t\}$ be admissible with respect to s in Part 1 of Proposition B.2.33 cannot be dropped.

Example B.2.3. Let α be a type, $p : \alpha \to o$, $C : \alpha$, $s \triangleq \Box_i(p\ x)$, and $t \triangleq C$. Also let I be an interpretation with two worlds w and w', an accessibility relation R_i such that $w\ R_i\ w'$, domain $\mathfrak{D}_\alpha \triangleq \{a, b\}$, and mapping V, where $V(p, w) = V(p, w') =$ relation that is true on a and false on b, and $V(C, w) = a$ and $V(C, w') = b$. Finally, let ν be any variable

assignment that maps x to a. Since C is not rigid, $\{x/C\}$ is not admissible with respect to $\Box_i(p\ x)$. Then

$$\begin{aligned}
&\mathcal{V}((\lambda x.\Box_i(p\ x)\ C), I, w, \nu)\\
&= \mathcal{V}(\lambda x.\Box_i(p\ x), I, w, \nu)(\mathcal{V}(C, I, w, \nu))\\
&= \mathcal{V}(\Box_i(p\ x), I, w, \nu'), \text{where } \nu' \text{ is } \nu \text{ except that } \nu'(x) = \mathcal{V}(C, I, w, \nu)\\
&= \mathcal{V}((p\ x), I, w', \nu'), \text{where } \nu' \text{ is } \nu \text{ except that } \nu'(x) = \mathcal{V}(C, I, w, \nu)\\
&= \mathsf{T}.
\end{aligned}$$

But

$$\begin{aligned}
&\mathcal{V}(\Box_i(p\ x)\{x/C\}, I, w, \nu)\\
&= \mathcal{V}(\Box_i(p\ C), I, w, \nu)\\
&= \mathcal{V}((p\ C), I, w', \nu)\\
&= \mathsf{F}.
\end{aligned}$$

B.2.7 Validity and Consequence

Now come the important concepts of validity and consequence.

Definition B.2.17. A biterm φ of rank n is *valid at a world w in an interpretation I* if $\mathcal{V}(\varphi, I, w, \nu) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}$, for every variable assignment ν.

Definition B.2.18. A biterm φ of rank n is *satisfiable at a world w in an interpretation I* if
$\mathcal{V}(\varphi, I, w, \nu) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}$, for some variable assignment ν.

The concepts of validity and satisfiability are dual to one another, as the next result shows.

Proposition B.2.34. *Let I be an interpretation and φ a biterm of rank n. Then the following hold.*

1. *φ is valid at w in I iff $\neg\varphi$ is not satisfiable at w in I.*
2. *φ is satisfiable at w in I iff $\neg\varphi$ is not valid at w in I.*

Proof. 1.

φ is valid at w in I
iff $\mathcal{V}(\varphi, I, w, \nu) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}$, for every variable assignment ν
iff $\mathcal{V}(\neg\varphi, I, w, \nu) = \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}$, for every variable assignment ν
iff $\neg\varphi$ is not satisfiable at w in I.

2.

φ is satisfiable at w in I
iff $\mathcal{V}(\varphi, I, w, \nu) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{F}$, for some variable assignment ν
iff $\mathcal{V}(\neg\varphi, I, w, \nu) \neq \lambda x_1.\ \ldots\ \lambda x_n.\mathsf{T}$, for some variable assignment ν
iff $\neg\varphi$ is not valid at w in I.

□

Definition B.2.19. A biterm φ is *valid in an interpretation* if it is valid at each world in the interpretation.

Proposition B.2.35. *Let φ be a biterm, I an interpretation, and w a world in I. Then the following hold.*

1. *φ is valid at w in I iff $\lambda x.\varphi$ is valid at w in I.*
2. *φ is valid in I iff $\lambda x.\varphi$ is valid in I.*
3. *φ is valid at w in I iff $(\Pi\ \varphi)$ is valid at w in I.*
4. *φ is valid in I iff $(\Pi\ \varphi)$ is valid in I.*
5. *φ is valid at w in I iff $\forall(\varphi)$ is valid at w in I.*
6. *φ is valid in I iff $\forall(\varphi)$ is valid in I.*

Proof. 1. This follows from Part 1 of Proposition B.2.4.
2. This follows from Part 1 and the definition of validity in an interpretation.
3. This follows from Part 2 of Proposition B.2.2.
4. This follows from Part 1 and the definition of validity in an interpretation.
5. This follows from Part 3 of Proposition B.2.4.
6. This follows from Part 3 and the definition of validity in an interpretation. □

Definition B.2.20. An interpretation $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}},\ V\rangle$ is said to be *based on* the frame $\langle W, \{R_i\}_{i=1}^m\rangle$.

Definition B.2.21. A biterm is *valid in a frame* if it is valid in every interpretation based on the frame.

Definition B.2.22. If L is a class of frames, then a biterm is L-*valid* if it is valid in each frame in L.

A theory is just a set of biterms, but it will be useful to distinguish two different kinds of biterms in the theory. This leads to the next definition.

Definition B.2.23. A *theory* is a pair $(\mathcal{G}, \mathcal{L})$, where $\mathcal{G}$ is a set of biterms, called *global assumptions*, and $\mathcal{L}$ is a set of biterms, called *local assumptions*.

Next comes the fundamental concept of a biterm being a consequence of a theory.

Definition B.2.24. Let L be a class of frames, $\mathcal{T} \triangleq (\mathcal{G}, \mathcal{L})$ a theory, and φ a biterm. Then φ is an L-*consequence* of the theory $\mathcal{T}$ if the following holds: for each interpretation I based on a frame in L for which all members of $\mathcal{G}$ are valid in I and for each world w in I such that each member of $\mathcal{L}$ is valid at w in I, it follows that φ is valid at w in I.

Notation. Let K denote the class of all frames. In the following, 'K-valid' is abbreviated to 'valid' and 'K-consequence' to 'consequence'.

For any class L of frames, $\mathtt{L} \subseteq \mathtt{K}$. Thus φ is a consequence of $\mathcal{T}$ implies that φ is an L-consequence of $\mathcal{T}$. Similarly, φ is valid implies that φ is L-valid.

Clearly each global assumption and each local assumption is a consequence of the theory. Hence, for any class L of frames, each global assumption and each local assumption is an L-consequence of the theory.

In applications, the concept of L-consequence is used in the following way. Each application has a distinguished pointed interpretation (I, w) known as the *intended pointed interpretation*, where I is based on a frame in some class L of frames. This means that, in the application, w is the actual world and I provides the worlds accessible to w by the various accessibility relations. Typically, in applications, given some term t, one needs to know the meaning of t in the intended pointed interpretation. If a formal definition of the intended pointed interpretation is available (as there is in the model checking setting), then this problem can be solved (under some finiteness assumptions). However, it is much more usual not to have a formal definition of the intended pointed interpretation, but instead to have a theory $\mathcal{T} \triangleq (\mathcal{G}, \mathcal{L})$ for the application. The assumption is that (I, w) and $\mathcal{T}$ are related in the following way:

- Each $\psi \in \mathcal{G}$ is valid in I.
- Each $\psi \in \mathcal{L}$ is valid at w in I.

Suppose now that a biterm φ is shown to be an L-consequence of $\mathcal{T}$. Then φ is valid at w in I. Intuitively, φ is 'true in the intended pointed interpretation'.

To emphasize, the key concept is that of meaning in the intended pointed interpretation, since it describes the actual situation the application has to deal with. But usually it is not possible for an application to have much detail about the intended pointed interpretation. Instead, it is usually much more feasible to be able to maintain a theory for the application, which serves as a *surrogate* for the intended pointed interpretation. The theory can then be used in many cases to answer the original question about the meaning of a term in the intended pointed interpretation. But note that this approach does not always work. For example, a biterm being true in the intended pointed interpretation is not the same as the biterm being a consequence of the theory: the latter also implies truth in any pointed interpretation satisfying the conditions in the preceding paragraph, not just the intended one, and usually there are many of these. Notwithstanding this difficulty, the convenience of working with the theory makes it worthwhile. Besides, in practice, there is usually little choice in the matter!

Sometimes it is convenient to have available the concept of a pointed model of a theory.

Definition B.2.25. Let $\mathcal{T} \triangleq (\mathcal{G}, \mathcal{L})$ be a theory. A *pointed model* for $\mathcal{T}$ is a pair (I, w) consisting of an interpretation I and a world w in I such that each $\psi \in \mathcal{G}$ is valid in I and each $\psi \in \mathcal{L}$ is valid at w in I.

In this terminology, a biterm φ is a consequence of $\mathcal{T}$ if, for each pointed model (I, w) of $\mathcal{T}$, φ is valid at w in I. Also, of course, an intended pointed interpretation should be a pointed model of the relevant theory.

The next result shows that it makes no difference semantically whether one works with the (possibly open) assumptions of a theory or the universal closures of all the assumptions (or, for that matter, anything in between).

Proposition B.2.36. *Let $\mathcal{T}$ be a theory and $\overline{\mathcal{T}}$ the theory obtained from $\mathcal{T}$ by taking as its global assumptions the universal closures of the global assumptions in $\mathcal{T}$ and its local assumptions the universal closures of the local assumptions in $\mathcal{T}$. Let* L *be a class of frames and φ a biterm. Then φ is an* L*-consequence of $\mathcal{T}$ iff φ is an* L*-consequence of $\overline{\mathcal{T}}$.*

Proof. The result follows immediately from Proposition B.2.35 and the definition of L-consequence. □

Note. Use will often be made of Proposition B.2.36 in the following by leaving off outer universal quantifiers in assumptions, thus avoiding syntactic clutter. In particular, this will usually be done with assumptions, such as those in the standard equality theory of Section B.3.1 and the definitions of list-processing functions, that are used in computations.

The next four results will be needed to show the correctness of the computational mechanism introduced in Appendix B.3.

Proposition B.2.37. *Let $\mathcal{T}$ be a theory and φ a global assumption of $\mathcal{T}$. Then, for each class* L *of frames and for each sequence $l_1 \dots l_p$ of indices, $\Box_{l_1} \cdots \Box_{l_p}\varphi$ in an* L*-consequence of $\mathcal{T}$.*

Proof. Suppose that $\mathcal{T}$ is $(\mathcal{G}, \mathcal{L})$. Let I be an interpretation based on a frame in L for which all members of $\mathcal{G}$ are valid in I and consider a world w in I such that each member of $\mathcal{L}$ is valid at w in I. Since φ is a global assumption, it is valid in I and therefore valid at each world v in I. Thus, whatever the sequence $l_1 \dots l_p$ of indices, $\Box_{l_1} \cdots \Box_{l_p}\varphi$ is valid at w in I. It follows that $\Box_{l_1} \cdots \Box_{l_p}\varphi$ in an L-consequence of $\mathcal{T}$. □

Proposition B.2.38. *Let $\mathcal{T}$ be a theory,* L *a class of frames, t a biterm, and $i \in \{1, \dots, m\}$. Then $\forall x.\Box_i t$ is an* L*-consequence of $\mathcal{T}$ iff $\Box_i \forall x.t$ is an* L*-consequence of $\mathcal{T}$.*

Proof. Let $\mathcal{T}$ be $(\mathcal{G}, \mathcal{L})$, I an interpretation based on a frame in L, and w a world in I. Suppose that each member of $\mathcal{G}$ is valid in I and that each member of $\mathcal{L}$ is valid at w in I. Then, by Part 1 of Proposition B.2.9, $\mathcal{V}(\forall x.\Box_i t, I, w, \nu) = \mathcal{V}(\Box_i \forall x.t, I, w, \nu)$, for each variable assignment ν. Hence $\forall x.\Box_i t$ is an L-consequence of $\mathcal{T}$ iff $\Box_i \forall x.t$ is an L-consequence of $\mathcal{T}$. □

Proposition B.2.39. *Let $\mathcal{T}$ be a theory,* L *a class of frames, and φ a biterm. Then φ is an* L*-consequence of $\mathcal{T}$ iff $\forall(\varphi)$ is an* L*-consequence of $\mathcal{T}$.*

Proof. Let $\mathcal{T}$ be $(\mathcal{G}, \mathcal{L})$, I an interpretation based on a frame in L, and w a world in I. Suppose that each member of $\mathcal{G}$ is valid in I and that each member of $\mathcal{L}$ is valid at w in I. Then, by Part 5 of Proposition B.2.35, φ is valid at w in I iff $\forall(\varphi)$ is valid at w in I. Hence φ is an L-consequence of $\mathcal{T}$ iff $\forall(\varphi)$ is an L-consequence of $\mathcal{T}$. □

Proposition B.2.40. *Let $\mathcal{T}$ be a theory,* L *a class of frames, $\Box_{l_1} \cdots \Box_{l_p}\varphi$ a biterm that is an* L*-consequence of $\mathcal{T}$, and θ a substitution that is admissible with respect to φ. Then $\Box_{l_1} \cdots \Box_{l_p}\varphi\theta$ is an* L*-consequence of $\mathcal{T}$.*

Proof. Let $\mathcal{T}$ be $(\mathcal{G}, \mathcal{L})$, I an interpretation based on a frame in L, and w a world in I. Suppose that each member of $\mathcal{G}$ is valid in I and that each member of $\mathcal{L}$ is valid at w in I. Since $\Box_{l_1} \cdots \Box_{l_p}\varphi$ is an L-consequence of $\mathcal{T}$, it follows that $\mathcal{V}(\Box_{l_1} \cdots \Box_{l_p}\varphi, I, w, \nu) =$

$\lambda x_1. \ldots \lambda x_n.\mathsf{T}$, for each variable assignment ν. By Proposition B.2.30, it follows that $\mathcal{V}(\Box_{l_1} \cdots \Box_{l_p} \varphi\theta, I, w, \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$, for each variable assignment ν. Thus $\Box_{l_1} \cdots \Box_{l_p} \varphi\theta$ is an $\mathtt{L}$-consequence of $\mathcal{T}$. □

Here are several kinds of relation that will be of interest.

Definition B.2.26. Let W be a non-empty set and R a binary relation on W.
R is *reflexive* if, for all $w \in W$, $w\,R\,w$.
R is *symmetric* if, for all $w, w' \in W$, $w\,R\,w'$ implies $w'\,R\,w$.
R is *transitive* if, for all $w, w', w'' \in W$, $w\,R\,w'$ and $w'\,R\,w''$ implies $w\,R\,w''$.
R is *serial* if, for all $w \in W$, there exists $w' \in W$ such that $w\,R\,w'$.
R is *Euclidean* if, for all $w, w', w'' \in W$, $w\,R\,w'$ and $w\,R\,w''$ implies $w'\,R\,w''$.

The next result gives logical characterizations of the various kinds of relation.

Proposition B.2.41. *Suppose that the alphabet contains a non-rigid constant p of type o, $\langle W, \{R_i\}_{i=1}^m\rangle$ is a frame, and $i \in \{1, \ldots, m\}$. Then the following hold.*

1. *R_i is reflexive iff $\Box_i\varphi \longrightarrow \varphi$ is valid in I, for every biterm φ and every interpretation I based on $\langle W, \{R_i\}_{i=1}^m\rangle$.*

2. *R_i is symmetric iff $\Diamond_i\Box_i\varphi \longrightarrow \varphi$ is valid in I, for every biterm φ and every interpretation I based on $\langle W, \{R_i\}_{i=1}^m\rangle$.*

3. *R_i is transitive iff $\Diamond_i\Diamond_i\varphi \longrightarrow \Diamond_i\varphi$ is valid in I, for every biterm φ and every interpretation I based on $\langle W, \{R_i\}_{i=1}^m\rangle$.*

4. *R_i is serial iff $\Box_i\varphi \longrightarrow \Diamond_i\varphi$ is valid in I, for every biterm φ and every interpretation I based on $\langle W, \{R_i\}_{i=1}^m\rangle$.*

5. *R_i is Euclidean iff $\Diamond_i\Box_i\varphi \longrightarrow \Box_i\varphi$ is valid in I, for every biterm φ and every interpretation I based on $\langle W, \{R_i\}_{i=1}^m\rangle$.*

Proof. 1. Suppose that R_i is reflexive. Let φ be a biterm having type $\alpha_1 \to \cdots \to \alpha_n \to o$, I an interpretation $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ based on $\langle W, \{R_i\}_{i=1}^m\rangle$, $w \in W$, and $d_i \in \mathcal{D}_{\alpha_i}$, for $i = 1, \ldots, n$. Suppose that $\mathcal{V}(\Box_i\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T}$, for every variable asignment ν. According to Part 1 of Proposition B.2.6, for each w' such that $w\,R_i\,w'$, $\mathcal{V}(\varphi, I, w', \nu)\ d_1 \ldots d_n = \mathsf{T}$, for every variable asignment ν. Since R_i is reflexive, $\mathcal{V}(\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T}$, for every variable asignment ν. By Part 3 of Proposition B.2.1, it follows that $\mathcal{V}(\Box_i\varphi \longrightarrow \varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T}$, for every variable asignment ν. That is, $\Box_i\varphi \longrightarrow \varphi$ is valid in I.

For the converse, suppose that R_i is not reflexive. Thus there exists $w \in W$ such that $w \not R_i\ w$. Let φ be the formula p and $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ any interpretation such that $V(p, v) = \mathsf{T}$, for all $v \neq w$, and $V(p, w) = \mathsf{F}$. Then $\mathcal{V}(\Box_i p \longrightarrow p, I, w, \nu) = \mathsf{F}$, for every variable assignment ν, and so $\Box_i p \longrightarrow p$ is not valid in I.

2. Suppose that R_i is symmetric. Let φ be a biterm having type $\alpha_1 \to \cdots \to \alpha_n \to o$, I an interpretation $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}}, V\rangle$ based on $\langle W, \{R_i\}_{i=1}^m\rangle$, $w \in W$, and $d_i \in \mathcal{D}_{\alpha_i}$, for $i = 1, \ldots, n$. Suppose that $\mathcal{V}(\Diamond_i\Box_i\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T}$, for every variable asignment ν. According to Part 2 of Proposition B.2.6, for some w' such that $w\,R_i\,w'$, $\mathcal{V}(\Box_i\varphi, I, w', \nu)\ d_1 \ldots d_n = \mathsf{T}$, for every variable asignment ν. Then, according to Part 1

of Proposition B.2.6, for each w'' such that $w' \, R_i \, w''$, $\mathcal{V}(\varphi, I, w'', \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. Since R_i is symmetric, $w' \, R_i \, w$ and so $\mathcal{V}(\varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. By Part 3 of Proposition B.2.1, it follows that $\mathcal{V}(\Diamond_i \Box_i \varphi \longrightarrow \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. That is, $\Diamond_i \Box_i \varphi \longrightarrow \varphi$ is valid in I.

For the converse, suppose that R_i is not symmetric. Thus there exists $w, w' \in W$ such that $w \, R_i \, w'$ but $w' \not R_i \, w$. Let φ be the formula p and $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, \; V \rangle$ any interpretation such that $V(p, v) = \mathsf{T}$ iff $w' \, R_i \, v$, for all $v \in W$. Then $\mathcal{V}(\Diamond_i \Box_i p, I, w, \nu) = \mathsf{T}$ and $\mathcal{V}(p, I, w, \nu) = \mathsf{F}$, for every variable assignment ν, and so $\Diamond_i \Box_i p \longrightarrow p$ is not valid in I.

3. Suppose that R_i is transitive. Let φ be a biterm having type $\alpha_1 \to \cdots \to \alpha_n \to o$, I an interpretation $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, \; V \rangle$ based on $\langle W, \{R_i\}_{i=1}^m \rangle$, $w \in W$, and $d_i \in \mathcal{D}_{\alpha_i}$, for $i = 1, \ldots, n$. Suppose that $\mathcal{V}(\Diamond_i \Diamond_i \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. According to Part 2 of Proposition B.2.6, for some w' such that $w \, R_i \, w'$, $\mathcal{V}(\Diamond_i \varphi, I, w', \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. Then, according to Part 2 of Proposition B.2.6 again, for some w'' such that $w' \, R_i \, w''$, $\mathcal{V}(\varphi, I, w'', \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. Since R_i is transitive, $w \, R_i \, w''$ and so $\mathcal{V}(\Diamond_i \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. By Part 3 of Proposition B.2.1, it follows that $\mathcal{V}(\Diamond_i \Diamond_i \varphi \longrightarrow \Diamond_i \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. That is, $\Diamond_i \Diamond_i \varphi \longrightarrow \Diamond_i \varphi$ is valid in I.

For the converse, suppose that R_i is not transitive. Thus there exist $w, w', w'' \in W$ such that $w \, R_i \, w'$ and $w' \, R_i \, w''$, but $w \not R_i \, w''$. Let φ be the formula p and $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, \; V \rangle$ any interpretation such that $V(p, v) = \mathsf{T}$ iff $w \not R_i \, v$, for all $v \in W$. Then $\mathcal{V}(\Diamond_i \Diamond_i p, I, w, \nu) = \mathsf{T}$ and $\mathcal{V}(\Diamond_i p, I, w, \nu) = \mathsf{F}$, for every variable assignment ν, and so $\Diamond_i \Diamond_i p \longrightarrow \Diamond_i p$ is not valid in I.

4. Suppose that R_i is serial. Let φ be a biterm having type $\alpha_1 \to \cdots \to \alpha_n \to o$, I an interpretation $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, \; V \rangle$ based on $\langle W, \{R_i\}_{i=1}^m \rangle$, $w \in W$, and $d_i \in \mathcal{D}_{\alpha_i}$, for $i = 1, \ldots, n$. Suppose that $\mathcal{V}(\Box_i \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. According to Part 1 of Proposition B.2.6, for each w' such that $w \, R_i \, w'$, $\mathcal{V}(\Diamond_i \varphi, I, w', \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. Since R_i is serial, there exists w' such that $w \, R_i \, w'$ and so $\mathcal{V}(\Diamond_i \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. By Part 3 of Proposition B.2.1, it follows that $\mathcal{V}(\Box_i \varphi \longrightarrow \Diamond_i \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. That is, $\Box_i \varphi \longrightarrow \Diamond_i \varphi$ is valid in I.

For the converse, suppose that R_i is not serial. Thus there exists $w \in W$ such that $w \not R_i \, w'$, for all $w' \in W$. Let φ be the formula p and $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, \; V \rangle$ any interpretation. Then $\mathcal{V}(\Box_i p, I, w, \nu) = \mathsf{T}$ and $\mathcal{V}(\Diamond_i p, I, w, \nu) = \mathsf{F}$, for every variable assignment ν, and so $\Box_i p \longrightarrow \Diamond_i p$ is not valid in I.

5. Suppose that R_i is Euclidean. Let φ be a biterm having type $\alpha_1 \to \cdots \to \alpha_n \to o$, I an interpretation $\langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, \; V \rangle$ based on $\langle W, \{R_i\}_{i=1}^m \rangle$, $w \in W$, and $d_i \in \mathcal{D}_{\alpha_i}$, for $i = 1, \ldots, n$. Suppose that $\mathcal{V}(\Diamond_i \Box_i \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. According to Part 2 of Proposition B.2.6, for some w'' such that $w \, R_i \, w''$, $\mathcal{V}(\Box_i \varphi, I, w'', \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. Then, according to Part 1 of Proposition B.2.6, for each w''' such that $w'' \, R_i \, w'''$, $\mathcal{V}(\varphi, I, w''', \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. Now consider w' such that $w \, R_i \, w'$. Since R_i is Euclidean, $w'' \, R_i \, w'$ and so $\mathcal{V}(\varphi, I, w', \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. Thus $\mathcal{V}(\Box_i \varphi, I, w, \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$, for every variable asignment ν. By Part 3 of Proposi-

tion B.2.1, it follows that $\mathcal{V}(\Diamond_i\Box_i\varphi \longrightarrow \Box_i\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{T}$, for every variable asignment ν. That is, $\Diamond_i\Box_i\varphi \longrightarrow \Box_i\varphi$ is valid in I.

For the converse, suppose that R_i is not Euclidean. Hence there exist $w, w', w'' \in W$ such that $w\,R_i\,w'$ and $w\,R_i\,w''$, but $w'\not R_i\,w''$. Let φ be the formula p and $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha\in\mathfrak{S}},\ V\rangle$ any interpretation such that $V(p, v) = \mathsf{T}$ iff $w'\,R_i\,v$, for all $v \in W$. Then $\mathcal{V}(\Diamond_i\Box_i p, I, w, \nu) = \mathsf{T}$, but $\mathcal{V}(\Box_i p, I, w, \nu) = \mathsf{F}$, for every variable assignment ν, and so $\Diamond_i\Box_i p \longrightarrow \Box_i p$ is not valid in I. □

To connect Proposition B.2.41 with the discussion about knowledge and belief in Section 5.1, note that

$\Diamond_i\Box_i\varphi \longrightarrow \varphi$ in Part 2 can be replaced by $\varphi \longrightarrow \Box_i\Diamond_i\varphi$

$\Diamond_i\Diamond_i\varphi \longrightarrow \Diamond_i\varphi$ in Part 3 can be replaced by $\Box_i\varphi \longrightarrow \Box_i\Box_i\varphi$

and

$\Diamond_i\Box_i\varphi \longrightarrow \Box_i\varphi$ in Part 5 can be replaced by $\Diamond_i\varphi \longrightarrow \Box_i\Diamond_i\varphi$.

The existence of non-rigid constants is needed in Proposition B.2.41. Instead of the assumption that there is a non-rigid propositional constant p, it is sufficient for there to exist two non-rigid constants a and b of the same type and p to be replaced by $a = b$. Alternatively, the propositional constant p could be replaced by a non-rigid biterm constant of any rank.

B.3 Reasoning

This section contains an account of a reasoning system that combines computation and proof, in the sense that a computation can call for a proof to be performed and, conversely, a proof can call for a computation to be performed. Thus the definitions below of computation and proof are mutually recursive. To get started, first the cases of pure computation, where no proof is involved, and pure proof, where no computation is involved, are considered, then later the two are put together.

B.3.1 Computation

This section studies the case of (pure) computation.

Consider the problem of determining the meaning of a term t in the intended pointed interpretation. If a formal definition of the intended pointed interpretation is available, then this problem can be solved (under some finiteness assumptions). However, it is assumed here that the intended pointed interpretation is not available, as is usually the case, so that the problem cannot be solved directly. Nevertheless, there is still a lot that can be done if the theory $\mathcal{T}$ of the application is available and enough of it is in equational form. Intuitively, if t can be 'simplified' sufficiently using $\mathcal{T}$, its meaning may become apparent even in the absence of detailed knowledge of the intended pointed interpretation. For example, if t can be simplified to a term containing only data constructors, then the meaning of t will be known since the data constructors have a fixed meaning in every interpretation.

More formally, the *computation problem* is as follows.

> Given a theory $\mathcal{T}$, a term t, and a sequence $\Box_{j_1} \cdots \Box_{j_r}$ of modalities, find a 'simpler' term t' such that $\Box_{j_1} \cdots \Box_{j_r} \forall (t = t')$ is a consequence of $\mathcal{T}$.

Thus t and t' have the same meaning in all worlds accessible from the point world in the intended pointed interpretation according to the modalities $\Box_{j_1} \cdots \Box_{j_r}$.

Here are the details about a mechanism that addresses the computational problem by employing equational reasoning to rewrite terms to 'simpler' terms that have the same meaning.

Definition B.3.1. Let $\mathcal{T} \triangleq (\mathcal{G}, \mathcal{L})$ be a theory. A *computation of rank 0 using* $\Box_{j_1} \cdots \Box_{j_r}$ *with respect to* $\mathcal{T}$ is a sequence $\{t_i\}_{i=1}^{n}$ of terms such that the following conditions are satisfied.

1. For $i = 1, \ldots, n-1$, there is
   (a) a subterm s_i of t_i at occurrence o_i, where the modal path to o_i in t_i is $k_1 \ldots k_{m_i}$,
   (b) i. a formula $\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_i}} \forall (u_i = v_i)$ in $\mathcal{L}$, or
   ii. a formula $\forall (u_i = v_i)$ in $\mathcal{G}$, and
   (c) a substitution θ_i that is admissible with respect to $u_i = v_i$

   such that $u_i\theta_i$ is α-equivalent to s_i and t_{i+1} is $t_i[s_i/v_i\theta_i]_{o_i}$.

The term t_1 is called the *goal* of the computation and t_n is called the *answer*.

Each subterm s_i is called a *redex*.

Each formula $\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_i}} \forall (u_i = v_i)$ or $\forall (u_i = v_i)$ is called an *input equation*.

The formula $\Box_{j_1} \cdots \Box_{j_r} \forall (t_1 = t_n)$ is called the *result* of the computation.

Note that, by Proposition B.1.9 and the remarks following Definition B.1.21, $t_i[s_i/v_i\theta_i]_{o_i}$ is a term of the same type as t_i, for $i = 1, \ldots, n$.

A *selection rule* chooses the redex at each step of a computation. A common selection rule is the *leftmost* one which chooses the leftmost outermost subterm that satisfies the requirements of Definition B.3.1. Note that it is straightforward to extend the definition of computation so that multiple redexes can be selected at each step. Then a common selection rule is the *parallel-outermost* one that selects all outermost subterms that each satisfy the requirements of Definition B.3.1. Later, especially for partial evaluation, there will be a need for selection rules that do not select the leftmost redex.

Computations generally require use of definitions of $=$, the connectives and quantifiers, and some other basic constants. These definitions, which constitute what is called the standard equality theory, are discussed now. But, before starting on this, note that since modalities are allowed one must distinguish between local and global assumptions in theories. Given the intended meanings of equality, the connectives and the quantifiers, it is natural that their definitions would normally be taken to be *global* assumptions in the theories of applications.

Now a series of definitions of $=$, the connectives, the constants Σ and Π, and so on, are given that constitute the standard equality theory. All substitutions appearing in these definitions are assumed to be admissible. Furthermore, some of the equations that follow are schemes which contain syntactical variables that are denoted by boldface

style. Syntactical variables are needed in situations where free variable capture is actually intended (and, therefore, the usual substitution mechanism is not appropriate). In general, a scheme is intended to stand for the collection of terms that can be obtained from the scheme by replacing its syntactical variables by terms, in such a way as to produce well-typed terms, of course. In a few places below, further *ad hoc* restrictions on the syntactical variables will be employed (such as requiring a syntactical variable to range over data constructors or object-level variables or rigid terms only). When using a scheme in a computation or proof, a choice of terms to replace its syntactical variables is first made. This results in a formula that can then be handled by Definitions B.3.1, B.3.4, B.3.6, and B.3.7. Thus each of the following schemes is a compact way of specifying a (possibly infinite) collection of formulas.

The first definition is that of the equality constant $=$. (Note that a % indicates that the remainder of the line is a comment.)

$= :\ a \to a \to o$

$(\boldsymbol{C}\ x_1 \ldots x_n = \boldsymbol{C}\ y_1 \ldots y_n) = (x_1 = y_1) \wedge \cdots \wedge (x_n = y_n)$

% where $\boldsymbol{C}$ is a data constructor of arity n.

$(\boldsymbol{C}\ x_1 \ldots x_n = \boldsymbol{D}\ y_1 \ldots y_m) = \bot$

% where $\boldsymbol{C}$ is a data constructor of arity n,

% $\boldsymbol{D}$ is a data constructor of arity m, and $\boldsymbol{C} \neq \boldsymbol{D}$.

$((x_1, \ldots, x_n) = (y_1, \ldots, y_n)) = (x_1 = y_1) \wedge \cdots \wedge (x_n = y_n)$

% where $n = 2, 3, \ldots$.

$(\lambda x.\boldsymbol{u} = \lambda y.\boldsymbol{v}) = (\mathit{less}\ \lambda x.\boldsymbol{u}\ \lambda y.\boldsymbol{v}) \wedge (\mathit{less}\ \lambda y.\boldsymbol{v}\ \lambda x.\boldsymbol{u})$

The first two schemes in the above definition simply capture the intended meaning of data constructors, while the third captures an important property of tuples. (Note that for the first scheme, if $n = 0$, then the body is $\top$.)

However, the fourth scheme is more subtle. In the proof-theoretic component below of the reasoning system, as is common in formulations of higher-order logics, one can adopt the axiom of extensionality:

$$\forall f.\forall g.((f = g) = \forall x.((f\ x) = (g\ x))).$$

However, this axiom is not used in the computational component of the reasoning system for the reason that it is not computationally useful: showing that $\forall x.((f\ x) = (g\ x))$ is not generally possible as there can be infinitely many values of x to consider. Instead, a special case of the axiom of extensionality is used in the standard equality theory. Its purpose is to provide a method of checking whether certain abstractions representing finite sets, finite multisets and similar data types are equal. In such cases, one can check for equality in a finite amount of time. The fourth scheme relies on the two following definitions.

$\mathit{less}\ :\ (a \to b) \to (a \to b) \to o$

$\mathit{less}\ \lambda x.\boldsymbol{d}\ z = \top$

% where $\boldsymbol{d}$ is a default term.

$$\mathit{less}\ (\lambda x.\mathit{if}\ \boldsymbol{u}\ \mathit{then}\ v\ \mathit{else}\ \boldsymbol{w})\ z =$$
$$(\forall x.(\boldsymbol{u} \longrightarrow v = (z\ x))) \wedge (\mathit{less}\ (\mathit{remove}\ \{x \mid \boldsymbol{u}\}\ \lambda x.\boldsymbol{w})\ z)$$

$$\mathit{remove}\ :\ (a \to o) \to (a \to b) \to (a \to b)$$

$$\mathit{remove}\ s\ \lambda x.\boldsymbol{d} = \lambda x.\boldsymbol{d}$$
% where $\boldsymbol{d}$ is a default term.
$$\mathit{remove}\ s\ \lambda x.\mathit{if}\ \boldsymbol{u}\ \mathit{then}\ v\ \mathit{else}\ \boldsymbol{w} =$$
$$\lambda x.\mathit{if}\ \boldsymbol{u} \wedge (x \notin s)\ \mathit{then}\ v\ \mathit{else}\ ((\mathit{remove}\ s\ \lambda x.\boldsymbol{w})\ x)$$

The intended meaning of *less* is best given by an illustration. Consider the multisets m and n. Then *less* m n is true iff each item in the support of m is also in the support of n and has the same multiplicity there. For sets, *less* is simply the subset relation. If s is a set and m a multiset, then *remove* s m returns the multiset obtained from m by removing all the items from its support that are in s.

The next definition is the obvious one for disequality.

$$\neq :\ a \to a \to o$$

$$x \neq y = \neg\ (x = y)$$

The following definitions are for the connectives $\wedge$, $\vee$, and $\neg$.

$$\wedge\ :\ o \to o \to o$$

$$\top \wedge x = x$$
$$x \wedge \top = x$$
$$\bot \wedge x = \bot$$
$$x \wedge \bot = \bot$$
$$(x \vee y) \wedge z = (x \wedge z) \vee (y \wedge z)$$
$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$$
$$(\mathit{if}\ u\ \mathit{then}\ v\ \mathit{else}\ w) \wedge t = \mathit{if}\ u \wedge t\ \mathit{then}\ v\ \mathit{else}\ w \wedge t$$
$$t \wedge (\mathit{if}\ u\ \mathit{then}\ v\ \mathit{else}\ w) = \mathit{if}\ t \wedge u\ \mathit{then}\ v\ \mathit{else}\ t \wedge w$$
$$u \wedge (\exists x_1.\cdots\exists x_n.\boldsymbol{v}) = \exists x_1.\cdots\exists x_n.(u \wedge \boldsymbol{v})$$
$$(\exists x_1.\cdots\exists x_n.\boldsymbol{u}) \wedge v = \exists x_1.\cdots\exists x_n.(\boldsymbol{u} \wedge v)$$
$$\boldsymbol{u} \wedge (\boldsymbol{x} = \boldsymbol{t}) \wedge \boldsymbol{v} = \boldsymbol{u}\{\boldsymbol{x}/\boldsymbol{t}\} \wedge (\boldsymbol{x} = \boldsymbol{t}) \wedge \boldsymbol{v}\{\boldsymbol{x}/\boldsymbol{t}\}$$
% where $\boldsymbol{x}$ is a variable free in $\boldsymbol{u}$ or $\boldsymbol{v}$ or both, but not free in $\boldsymbol{t}$,
% and $\boldsymbol{t}$ is not a variable.
$$\boldsymbol{u} \wedge (\boldsymbol{t} = \boldsymbol{x}) \wedge \boldsymbol{v} = \boldsymbol{u}\{\boldsymbol{x}/\boldsymbol{t}\} \wedge (\boldsymbol{x} = \boldsymbol{t}) \wedge \boldsymbol{v}\{\boldsymbol{x}/\boldsymbol{t}\}$$
% where $\boldsymbol{x}$ is a variable free in $\boldsymbol{u}$ or $\boldsymbol{v}$ or both, but not free in $\boldsymbol{t}$,
% and $\boldsymbol{t}$ is not a variable.

$$\vee\ :\ o \to o \to o$$

$$\top \vee x = \top \tag{O1}$$
$$x \vee \top = \top \tag{O2}$$
$$\bot \vee x = x$$
$$x \vee \bot = x$$
$$(\mathit{if}\ u\ \mathit{then}\ \top\ \mathit{else}\ w) \vee t = \mathit{if}\ u\ \mathit{then}\ \top\ \mathit{else}\ w \vee t$$
$$(\mathit{if}\ u\ \mathit{then}\ \bot\ \mathit{else}\ w) \vee t = (\neg\, u \wedge w) \vee t$$
$$t \vee (\mathit{if}\ u\ \mathit{then}\ \top\ \mathit{else}\ w) = \mathit{if}\ u\ \mathit{then}\ \top\ \mathit{else}\ t \vee w$$
$$t \vee (\mathit{if}\ u\ \mathit{then}\ \bot\ \mathit{else}\ w) = t \vee (\neg\, u \wedge w)$$

$$\neg\ :\ o \to o$$

$$\neg \bot = \top$$
$$\neg \top = \bot$$
$$\neg (\neg\, x) = x$$
$$\neg (x \wedge y) = (\neg\, x) \vee (\neg\, y)$$
$$\neg (x \vee y) = (\neg\, x) \wedge (\neg\, y)$$
$$\neg (\mathit{if}\ u\ \mathit{then}\ v\ \mathit{else}\ w) = \mathit{if}\ u\ \mathit{then} \neg\, v\ \mathit{else} \neg\, w$$

These definitions are straightforward, except perhaps for the last three schemes in the definition of $\wedge$. The second and third last schemes allow the scope of existential quantifiers to be extended provided it does not result in free variable capture. (Recall the convention restricting the possible term replacements for syntactical variables.)

The last scheme allows the elimination of some occurrences of a free variable ($\boldsymbol{x}$, in this case), thus simplifying an expression. A similar scheme allowing this kind of simplification also occurs in the definition of Σ below. However, a few words about the expression $\boldsymbol{u} \wedge (\boldsymbol{x} = \boldsymbol{t}) \wedge \boldsymbol{v}$ are necessary. The intended meaning of this expression is that it is a term such that $(\boldsymbol{x} = \boldsymbol{t})$ is 'embedded conjunctively' inside it. More formally, a term t is embedded conjunctively in t and, if t is embedded conjunctively in r (or s), then t is embedded conjunctively in $r \wedge s$. So, for example, $x = s$ is embedded conjunctively in the term $((p \wedge q) \wedge r) \wedge ((x = s) \wedge (t \wedge u))$.

Next come the definitions of Σ and Π.

$$\Sigma\ :\ (a \to o) \to o$$

$$\exists x_1. \cdots \exists x_n. \top = \top$$
$$\exists x_1. \cdots \exists x_n. \bot = \bot$$
$$\exists x_1. \cdots \exists x_n. (\boldsymbol{x} \wedge (x_i = \boldsymbol{u}) \wedge \boldsymbol{y}) =$$
$$\qquad \exists x_1. \cdots \exists x_{i-1}. \exists x_{i+1}. \cdots \exists x_n. (\boldsymbol{x}\{x_i/\boldsymbol{u}\} \wedge \boldsymbol{y}\{x_i/\boldsymbol{u}\})$$
$$\qquad \%\ \text{where } x_i \text{ is not free in } \boldsymbol{u}.$$
$$\exists x_1. \cdots \exists x_n. (\boldsymbol{x} \wedge (\boldsymbol{u} = x_i) \wedge \boldsymbol{y}) =$$
$$\qquad \exists x_1. \cdots \exists x_{i-1}. \exists x_{i+1}. \cdots \exists x_n. (\boldsymbol{x}\{x_i/\boldsymbol{u}\} \wedge \boldsymbol{y}\{x_i/\boldsymbol{u}\})$$
$$\qquad \%\ \text{where } x_i \text{ is not free in } \boldsymbol{u}.$$
$$\exists x_1. \cdots \exists x_n. (\boldsymbol{u} \vee \boldsymbol{v}) = (\exists x_1. \cdots \exists x_n. \boldsymbol{u}) \vee (\exists x_1. \cdots \exists x_n. \boldsymbol{v})$$

$\exists x_1.\cdots\exists x_n.(\mathit{if}\ \boldsymbol{u}\ \mathit{then}\ \top\ \mathit{else}\ \boldsymbol{w}) =$
$\quad \mathit{if}\ \exists x_1.\cdots\exists x_n.\boldsymbol{u}\ \mathit{then}\ \top\ \mathit{else}\ \exists x_1.\cdots\exists x_n.\boldsymbol{w}$
$\exists x_1.\cdots\exists x_n.(\mathit{if}\ \boldsymbol{u}\ \mathit{then}\ \bot\ \mathit{else}\ \boldsymbol{w}) = \exists x_1.\cdots\exists x_n.(\neg\ \boldsymbol{u} \wedge \boldsymbol{w})$
$\exists x_1.\cdots\exists x_n.(\mathit{if}\ \boldsymbol{u}\ \mathit{then}\ \boldsymbol{v}\ \mathit{else}\ \boldsymbol{w}) =$
$\quad \mathit{if}\ \exists x_1.\cdots\exists x_n.(\boldsymbol{u} \wedge \boldsymbol{v})\ \mathit{then}\ \top\ \mathit{else}\ \exists x_1.\cdots\exists x_n.(\neg\boldsymbol{u} \wedge \boldsymbol{w})$
$\quad$ % where $\boldsymbol{v}$ is neither $\top$ nor $\bot$.

$\Pi\ :\ (a \to o) \to o$

$\forall x_1.\cdots\forall x_n.(\bot \longrightarrow \boldsymbol{u}) = \top$
$\forall x_1.\cdots\forall x_n.(\boldsymbol{x} \wedge (x_i = \boldsymbol{u}) \wedge \boldsymbol{y} \longrightarrow \boldsymbol{v}) =$
$\quad \forall x_1.\cdots\forall x_{i-1}.\forall x_{i+1}.\cdots\forall x_n.(\boldsymbol{x}\{x_i/\boldsymbol{u}\} \wedge \boldsymbol{y}\{x_i/\boldsymbol{u}\} \longrightarrow \boldsymbol{v}\{x_i/\boldsymbol{u}\})$
$\quad$ % where x_i is not free in $\boldsymbol{u}$.
$\forall x_1.\cdots\forall x_n.(\boldsymbol{x} \wedge (\boldsymbol{u} = x_i) \wedge \boldsymbol{y} \longrightarrow \boldsymbol{v}) =$
$\quad \forall x_1.\cdots\forall x_{i-1}.\forall x_{i+1}.\cdots\forall x_n.(\boldsymbol{x}\{x_i/\boldsymbol{u}\} \wedge \boldsymbol{y}\{x_i/\boldsymbol{u}\} \longrightarrow \boldsymbol{v}\{x_i/\boldsymbol{u}\})$
$\quad$ % where x_i is not free in $\boldsymbol{u}$.
$\forall x_1.\cdots\forall x_n.(\boldsymbol{u} \vee \boldsymbol{v} \longrightarrow \boldsymbol{t}) =$
$\quad (\forall x_1.\cdots\forall x_n.(\boldsymbol{u} \longrightarrow \boldsymbol{t})) \wedge (\forall x_1.\cdots\forall x_n.(\boldsymbol{v} \longrightarrow \boldsymbol{t}))$
$\forall x_1.\cdots\forall x_n.((\mathit{if}\ \boldsymbol{u}\ \mathit{then}\ \top\ \mathit{else}\ \boldsymbol{v}) \longrightarrow \boldsymbol{t}) =$
$\quad (\forall x_1.\cdots\forall x_n.(\boldsymbol{u} \longrightarrow \boldsymbol{t})) \wedge (\forall x_1.\cdots\forall x_n.(\boldsymbol{v} \longrightarrow \boldsymbol{t}))$
$\forall x_1.\cdots\forall x_n.((\mathit{if}\ \boldsymbol{u}\ \mathit{then}\ \bot\ \mathit{else}\ \boldsymbol{v}) \longrightarrow \boldsymbol{t}) = \forall x_1.\cdots\forall x_n.(\neg\ \boldsymbol{u} \wedge \boldsymbol{v} \longrightarrow \boldsymbol{t})$

For the definition of Σ, recall that $\exists x.t$ stands for $(\Sigma\ \lambda x.t)$. Thus, for example, $\exists x.\top$ stands for $(\Sigma\ \lambda x.\top)$.

The definition for the *if_then_else* constant follows.

if_then_else $:\ o \times a \times a \to a$

$\mathit{if}\ \top\ \mathit{then}\ u\ \mathit{else}\ v = u$
$\mathit{if}\ \bot\ \mathit{then}\ u\ \mathit{else}\ v = v$

The next two assumptions involve function application and the *if_then_else* constant.

$(\boldsymbol{w}\ \ \mathit{if}\ \boldsymbol{x} = \boldsymbol{t}\ \mathit{then}\ u\ \mathit{else}\ v) = \mathit{if}\ \boldsymbol{x} = \boldsymbol{t}\ \mathit{then}\ (\boldsymbol{w}\{\boldsymbol{x}/\boldsymbol{t}\}\ u)\ \mathit{else}\ (\boldsymbol{w}\ v)$
$\quad$ % where $\boldsymbol{x}$ is a variable.
$(\mathit{if}\ \boldsymbol{x} = \boldsymbol{t}\ \mathit{then}\ u\ \mathit{else}\ v\ \ \boldsymbol{w}) = \mathit{if}\ \boldsymbol{x} = \boldsymbol{t}\ \mathit{then}\ (u\ \boldsymbol{w}\{\boldsymbol{x}/\boldsymbol{t}\})\ \mathit{else}\ (v\ \boldsymbol{w})$
$\quad$ % where $\boldsymbol{x}$ is a variable.

The following assumption provides β-reduction.

$(\lambda x.\boldsymbol{u}\ t) = \boldsymbol{u}\{x/t\}$

Finally, there are three schemes that involve modalities.

$$\Box_i \boldsymbol{t} = \boldsymbol{t} \tag{M}$$

% where $\boldsymbol{t}$ is a rigid term and $i \in \{1, \ldots, m\}$.

$$(\Box_i \mathbf{s}\ \mathbf{t}) = \Box_i(\mathbf{s}\ \mathbf{t})$$

% where $\mathbf{t}$ is a rigid term and $i \in \{1, \ldots, m\}$.

$$\Box_i \lambda x.\mathbf{t} = \lambda x.\Box_i \mathbf{t}$$

% where $i \in \{1, \ldots, m\}$.

Definition B.3.2. The *standard equality theory* is the theory consisting of the preceding definitions of equality and subsidiary constants, the connectives, the quantifiers, the *if_then_else* constant, β-reduction, and the three modal schemes as global assumptions (and no local assumptions).

The soundness of computation is given by the next result.

Proposition B.3.1. *Let $\mathcal{T}$ be a theory. Then the result of a computation of rank 0 using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{T}$ is a consequence of $\mathcal{T}$.*

Proof. The proof is by induction on the length n of computations. If $n = 1$, then the result is obvious.

Suppose now that the result holds for computations of length n. Consider a computation $\{t_i\}_{i=1}^{n+1}$ using $\Box_{j_1} \cdots \Box_{j_r}$ of length $n+1$ for which, at the nth step, the input equation is either the local assumption

$$\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_n}} \forall (u_n = v_n)$$

or the global assumption $\forall(u_n = v_n)$, the admissible substitution is θ_n, and the redex is s_n at occurrence o_n with modal path $k_1 \ldots k_{m_n}$ to o_n.

Let $\mathcal{T}$ be $(\mathcal{G}, \mathcal{L})$, I an interpretation, and w a world in I. Suppose that each member of $\mathcal{G}$ is valid in I and that each member of $\mathcal{L}$ is valid at w in I. By the induction hypothesis, $\Box_{j_1} \cdots \Box_{j_r} \forall(t_1 = t_n)$ is valid at w in I, that is, $\mathcal{V}(\Box_{j_1} \cdots \Box_{j_r} \forall(t_1 = t_n), I, w, \nu) = \mathsf{T}$, for each variable assignment ν. Thus, by Proposition B.2.7, $\mathcal{V}(\forall(t_1 = t_n), I, w', \nu) = \mathsf{T}$, for each variable assignment ν and each w' such that $w\ R_{j_1} \circ \cdots \circ R_{j_r}\ w'$. Hence, by Part 3 of Proposition B.2.4, $\mathcal{V}(t_1 = t_n, I, w', \nu) = \mathsf{T}$, for each variable assignment ν and each w' such that $w\ R_{j_1} \circ \cdots \circ R_{j_r}\ w'$.

By Proposition B.2.37, for either case of input equation,

$$\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_n}} \forall (u_n = v_n)$$

is a consequence of $\mathcal{T}$. Thus, by Proposition B.2.38,

$$\forall(\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_n}} (u_n = v_n))$$

is a consequence of $\mathcal{T}$ and so, by Proposition B.2.39,

$$\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_n}} (u_n = v_n)$$

is a consequence of $\mathcal{T}$. Hence, by Proposition B.2.40,

$$\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_n}} (u_n\theta_n = v_n\theta_n)$$

is a consequence of $\mathcal{T}$. Thus

$$\mathcal{V}(\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_n}} (u_n\theta_n = v_n\theta_n), I, w, \nu) = \mathsf{T},$$

for each variable assignment ν. Consequently, by Proposition B.2.7,

$$\mathcal{V}(\Box_{k_1} \cdots \Box_{k_{m_n}} (u_n\theta_n = v_n\theta_n), I, w', \nu) = \mathsf{T},$$

for each variable assignment ν and each w' such that $w\ R_{j_1} \circ \cdots \circ R_{j_r}\ w'$.

Now t_{n+1} is $t_n[s_n/v_n\theta_n]_{o_n}$. Hence, it follows from Proposition B.2.31 that $\mathcal{V}(t_n = t_{n+1}, I, w', \nu) = \mathsf{T}$, for each variable assignment ν and each w' such that $w\ R_{j_1} \circ \cdots \circ R_{j_r}\ w'$. By Proposition B.2.10, $\mathcal{V}(t_1 = t_{n+1}, I, w', \nu) = \mathsf{T}$, for each variable assignment ν and each w' such that $w\ R_{j_1} \circ \cdots \circ R_{j_r}\ w'$. By Part 3 of Proposition B.2.4, $\mathcal{V}(\forall(t_1 = t_{n+1}), I, w', \nu) = \mathsf{T}$, for each variable assignment ν and each w' such that $w\ R_{j_1} \circ \cdots \circ R_{j_r}\ w'$. By Proposition B.2.7, $\mathcal{V}(\Box_{j_1} \cdots \Box_{j_r} \forall(t_1 = t_{n+1}), I, w, \nu) = \mathsf{T}$, for each variable assignment ν. Thus $\Box_{j_1} \cdots \Box_{j_r} \forall(t_1 = t_{n+1})$ is a consequence of $\mathcal{T}$. This completes the induction step. □

The next proposition provides a useful result about computing with basic terms.

Proposition B.3.2. *Let s and t be basic terms of the same type, $\mathcal{E}$ the standard equality theory, and $\Box_{j_1} \cdots \Box_{j_r}$ any sequence of modalities. Then the following hold.*

1. *If $s \overset{*}{\longleftrightarrow}_\alpha t$, then there is a computation of rank 0 using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{E}$ for the goal $s = t$ that has the result $\Box_{j_1} \cdots \Box_{j_r}((s = t) = \top)$.*

2. *If $s \not\overset{*}{\longleftrightarrow}_\alpha t$, then there is a computation of rank 0 using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{E}$ for the goal $s = t$ that has the result $\Box_{j_1} \cdots \Box_{j_r}((s = t) = \bot)$.*

Proof. The proof of both parts together is by induction on the structure of basic terms [96, Proposition 3.5.1].

Suppose that s of type $T\ \alpha_1 \ldots \alpha_k$ has the form $C\ s_1 \ldots s_n$. If $s \overset{*}{\longleftrightarrow}_\alpha t$, then t has the form $C\ t_1 \ldots t_n$ and $s_i \overset{*}{\longleftrightarrow}_\alpha t_i$, for $i = 1, \ldots, n$. By the induction hypothesis, there is a computation using $\Box_{j_1} \cdots \Box_{j_r}$ with the goal $s_i = t_i$ and the answer $\top$, for $i = 1, \ldots, n$. Then, using the global assumption

$$(\boldsymbol{C}\ x_1 \ldots x_n = \boldsymbol{C}\ y_1 \ldots y_n) = (x_1 = y_1) \wedge \cdots \wedge (x_n = y_n)$$

and other global assumptions in $\mathcal{E}$, a computation can be constructed using $\Box_{j_1} \cdots \Box_{j_r}$ with the goal $s = t$ and the answer $\top$, as required.

If $s \not\overset{*}{\longleftrightarrow}_\alpha t$, then there are two cases to consider. Suppose first that t has the form $D\ t_1 \ldots t_m$, where $C \neq D$. Then the global assumption

$$(\boldsymbol{C}\ x_1 \ldots x_n = \boldsymbol{D}\ y_1 \ldots y_m) = \bot$$

can be used to construct a computation with the goal $s = t$ and the answer $\bot$. In the second case, t has the form $C\ t_1 \ldots t_n$ and $s_j \not\overset{*}{\longleftrightarrow}_\alpha t_j$, for some $j \in \{1, \ldots, n\}$. By the

induction hypothesis, there is a computation using $\Box_{j_1} \cdots \Box_{j_r}$ with the goal $s_j = t_j$ and the answer $\bot$. Then, using the global assumption

$$(\boldsymbol{C}\ x_1 \ldots x_n = \boldsymbol{C}\ y_1 \ldots y_n) = (x_1 = y_1) \wedge \cdots \wedge (x_n = y_n)$$

and other global assumptions in $\mathcal{E}$, a computation can be constructed using $\Box_{j_1} \cdots \Box_{j_r}$ with the goal $s = t$ and the answer $\bot$, as required. This completes the case for basic structures.

Suppose next that s of type $\gamma \to \eta$ has the form

$$\lambda x.\mathit{if}\ x = t_1\ \mathit{then}\ s_1\ \mathit{else}\ \ldots\ \mathit{if}\ x = t_n\ \mathit{then}\ s_n\ \mathit{else}\ s_0,$$

where $t_1, \ldots, t_n \in \mathfrak{B}_\gamma$, $s_1, \ldots, s_n \in \mathfrak{B}_\eta$, $t_1 < \cdots < t_n$, $s_i \notin \mathfrak{D}_\eta$, for $1 \leq i \leq n$ $(n \in \mathbb{N}_0)$, and $s_0 \in \mathfrak{D}_\eta$. There are two cases to consider.

The first case to consider is when $s \overset{*}{\longleftrightarrow}_\alpha t$. In this case, t must have the form

$$\lambda y.\mathit{if}\ y = t'_1\ \mathit{then}\ s'_1\ \mathit{else}\ \ldots\ \mathit{if}\ y = t'_n\ \mathit{then}\ s'_n\ \mathit{else}\ s_0,$$

where $t'_1, \ldots, t'_n \in \mathfrak{B}_\gamma$, $s'_1, \ldots, s'_n \in \mathfrak{B}_\eta$, $t'_1 < \cdots < t'_n$, $s_i \notin \mathfrak{D}_\eta$, for $1 \leq i \leq n$ $(n \in \mathbb{N}_0)$, $s_0 \in \mathfrak{D}_\eta$, and $s_i \overset{*}{\longleftrightarrow}_\alpha s'_i$ and $t_i \overset{*}{\longleftrightarrow}_\alpha t'_i$, for $i = 1, \ldots, n$. By the induction hypothesis, there is a computation using using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{E}$ for the goal $t_i = t'_i$ that has the result $\Box_{j_1} \cdots \Box_{j_r}((t_i = t'_i) = \top)$ and there is a computation using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{E}$ for the goal $s_i = s'_i$ that has the result $\Box_{j_1} \cdots \Box_{j_r}((s_i = s'_i) = \top)$, for $i = 1, \ldots, n$. A computation is now constructed for the goal $s = t$. In this computation and the later ones in the proof, the easily established fact that, if s and t are basic terms and $s < t$, then $s \not\overset{*}{\longleftrightarrow}_\alpha t$ is used several times. Let s_{rest} denote

$$\lambda x.\mathit{if}\ x = t_2\ \mathit{then}\ s_2\ \mathit{else}\ \ldots\ \mathit{if}\ x = t_n\ \mathit{then}\ s_n\ \mathit{else}\ s_0.$$

An outline of this computation is as follows. (The redexes are underlined.)

$$\begin{array}{l}
\underline{s = t} \\
\underline{(\mathit{less}\ s\ t)} \wedge (\mathit{less}\ t\ s) \\
(\underline{(\forall x.((x = t_1) \longrightarrow s_1 = (t\ x)))} \wedge \\
\qquad\qquad (\mathit{less}\ (\mathit{remove}\ \{x \mid (x = t_1)\}\ s_{rest})\ t)) \wedge (\mathit{less}\ t\ s) \\
((s_1 = \underline{(t\ t_1)}) \wedge (\mathit{less}\ (\mathit{remove}\ \{x \mid (x = t_1)\}\ s_{rest})\ t)) \wedge (\mathit{less}\ t\ s) \\
((s_1 = (\mathit{if}\ \underline{t_1 = t'_1}\ \mathit{then}\ s'_1\ \mathit{else}\ \ldots\ \mathit{if}\ t_1 = t'_n\ \mathit{then}\ s'_n\ \mathit{else}\ s_0)) \wedge \\
\qquad\qquad (\mathit{less}\ (\mathit{remove}\ \{x \mid (x = t_1)\}\ s_{rest})\ t)) \wedge (\mathit{less}\ t\ s) \\
\qquad \vdots \\
((s_1 = \underline{(\mathit{if}\ \top\ \mathit{then}\ s'_1\ \mathit{else}\ \ldots\ \mathit{if}\ t_1 = t'_n\ \mathit{then}\ s'_n\ \mathit{else}\ s_0)}) \wedge \\
\qquad\qquad (\mathit{less}\ (\mathit{remove}\ \{x \mid (x = t_1)\}\ s_{rest})\ t)) \wedge (\mathit{less}\ t\ s) \\
(\underline{(s_1 = s'_1)} \wedge (\mathit{less}\ (\mathit{remove}\ \{x \mid (x = t_1)\}\ s_{rest})\ t)) \wedge (\mathit{less}\ t\ s) \\
\qquad \vdots \\
\underline{(\top \wedge (\mathit{less}\ (\mathit{remove}\ \{x \mid (x = t_1)\}\ s_{rest})\ t))} \wedge (\mathit{less}\ t\ s)
\end{array}$$

$$(less\ \underline{(remove\ \{x \mid (x = t_1)\}\ s_{rest})}\ t) \wedge (less\ t\ s)$$

$$\vdots$$

$$\underline{(less\ s_{rest}\ t)} \wedge (less\ t\ s)$$

$$\vdots$$

$$\underline{\top \wedge (less\ t\ s)}$$

$$\underline{less\ t\ s}$$

$$\vdots$$

$$\top.$$

The case when $n = 0$, that is, when s and t have the form $\lambda x.s_0$, is much easier.

The second case to consider for basic abstractions is when $s \not\xleftrightarrow{*}_\alpha t$. In this case, there exists $j \in \{1, \ldots, n\}$ such that $t_i \xleftrightarrow{*}_\alpha t'_i$ and $s_i \xleftrightarrow{*}_\alpha s'_i$, for $i < j$, and either $t_j \not\xleftrightarrow{*}_\alpha t'_j$ or $s_j \not\xleftrightarrow{*}_\alpha s'_j$. Alternatively, one of s and t is a strict "prefix" (up to α-equivalence) of the other; the details of this case are omitted. Suppose that $t_j \xleftrightarrow{*}_\alpha t'_j$, but $s_j \not\xleftrightarrow{*}_\alpha s'_j$. Let s_{rest_j} denote

$$\lambda x.if\ x = t_j\ then\ s_j\ else\ \ldots\ if\ x = t_n\ then\ s_n\ else\ s_0.$$

An outline of a computation for the goal $s = t$ in this case is as follows.

$$\underline{s = t}$$

$$\underline{(less\ s\ t)} \wedge (less\ t\ s)$$

$$\vdots$$

$$\underline{(less\ s_{rest_j}\ t)} \wedge (less\ t\ s)$$

$$\underline{((\forall x.((x = t_j) \longrightarrow s_j = (t\ x)))} \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$((s_j = \underline{(t\ t_j)}) \wedge (less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$((s_j = (if\ \underline{t_j = t'_1}\ then\ s'_1\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0)) \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$((s_j = (\underline{if\ \bot\ then\ s'_1\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0})) \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$((s_j = (if\ \underline{t_j = t'_2}\ then\ s'_2\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0)) \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$((s_j = (if\ \underline{t_j = t'_j}\ then\ s'_j\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0)) \wedge$$

$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$((s_j = (\underline{if\ \top\ then\ s'_j\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0})) \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$(\underline{(s_j = s'_j)} \wedge (less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$\underline{(\bot \wedge (less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t))} \wedge (less\ t\ s)$$

$$\underline{\bot \wedge (less\ t\ s)}$$

$$\bot.$$

The remaining case is when $t_j \not\xrightarrow{*}_{\alpha} t'_j$. Suppose that $t_j < t'_j$. (The case when $t'_j < t_j$ is similar except that $(less\ t\ s)$ instead of $(less\ s\ t)$ is shown to evaluate to $\bot$.) An outline of a computation for the goal $s = t$ in this case is as follows.

$$\underline{s = t}$$

$$\underline{(less\ s\ t)} \wedge (less\ t\ s)$$

$$\vdots$$

$$\underline{(less\ s_{rest_j}\ t)} \wedge (less\ t\ s)$$

$$(\underline{(\forall x.((x = t_j) \longrightarrow s_j = (t\ x)))} \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$((s_j = \underline{(t\ t_j)}) \wedge (less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$((s_j = (\underline{if\ t_j = t'_1\ then\ s'_1\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0})) \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$((s_j = (\underline{if\ \bot\ then\ s'_1\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0})) \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$((s_j = (\underline{if\ t_j = t'_j\ then\ s'_j\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0})) \wedge$$
$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$((s_j = (\underline{if\ \bot\ then\ s'_j\ else\ \ldots\ if\ t_j = t'_n\ then\ s'_n\ else\ s_0})) \wedge$$

$$(less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$\underline{((s_j = s_0)} \wedge (less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t)) \wedge (less\ t\ s)$$

$$\vdots$$

$$\underline{(\bot \wedge (less\ (remove\ \{x \mid (x = t_j)\}\ s_{rest_j})\ t))} \wedge (less\ t\ s)$$

$$\underline{\bot \wedge (less\ t\ s)}$$

$$\bot.$$

This completes the case of basic abstractions.

Suppose finally that s of type $\sigma \triangleq (\sigma_1, \ldots, \sigma_n)$ has the form $(s_1, \ldots, s_n)$. Thus t has the form $(t_1, \ldots, t_n)$. If $s \overset{*}{\longleftrightarrow}_\alpha t$, then $s_i \overset{*}{\longleftrightarrow}_\alpha t_i$, for $i = 1, \ldots, n$. By the induction hypothesis, there is a computation using $\Box_{j_1} \cdots \Box_{j_r}$ with the goal $s_i = t_i$ and the answer $\top$, for $i = 1, \ldots, n$. Then, using the global assumption

$$((x_1, \ldots, x_n) = (y_1, \ldots, y_n)) = (x_1 = y_1) \wedge \cdots \wedge (x_n = y_n)$$

and other global assumptions in $\mathcal{E}$, a computation can be constructed using $\Box_{j_1} \cdots \Box_{j_r}$ with the goal $s = t$ and the answer $\top$, as required. If $s \not\overset{*}{\longleftrightarrow}_\alpha t$, then $s_j \not\overset{*}{\longleftrightarrow}_\alpha t_j$, for some $j \in \{1, \ldots, n\}$. By the induction hypothesis, there is a computation using $\Box_{j_1} \cdots \Box_{j_r}$ with the goal $s_j = t_j$ and the answer $\bot$. A similar argument to the previous case now shows that there is a computation using $\Box_{j_1} \cdots \Box_{j_r}$ with the goal $s = t$ and the answer $\bot$, as required. This completes the case of basic tuples. □

B.3.2 Proof

Next the discussion turns to (pure) proof using a tableau system.

Consider the problem of determining whether a biterm φ is true or false in the intended pointed interpretation. Once again suppose that a formal definition of the intended pointed interpretation is not available, but that the theory $\mathcal{T}$ of the application is available. Then the problem can still be solved by showing that φ is a theorem of $\mathcal{T}$. In this case, φ is a consequence of $\mathcal{T}$, by the soundness of the proof system, so that φ must be true in the intended pointed interpretation. Theoremhood is, of course, undecidable in the case of higher-order logic, but it is still possible to prove theorems in many cases of practical interest.

More formally, the *proof problem* is as follows.

> Given a theory $\mathcal{T}$ and biterm φ, determine whether φ is a consequence of $\mathcal{T}$.

Here now are the details of a proof system that can determine consequence.

The tableau system employs prefixed biterms as is often the case for modal logics.

Definition B.3.3. A *prefix* is a finite sequence of the form

$$1.\langle n_1, j_1\rangle. \ldots .\langle n_k, j_k\rangle,$$

where n_i is a positive integer and $j_i \in \{1, \ldots, m\}$, for $i = 1, \ldots, k$.

A *prefixed biterm* is an expression of the form $\sigma\ \varphi$, where σ is a prefix and φ is a biterm.

In the following, $\langle n, j\rangle$ is abbreviated to n_j.

The tableau system works as usual for modal logics. Given that the aim is to show that a biterm φ is an L-consequence of sets of global and local assumptions, the initial node of the tableau is $1\ \neg\varphi$. At any stage in the construction of the tableau, one of its branches is chosen and the branch extended by one of the tableau rules. In this book, there is a concentration on the (multi-modal) logic $\mathbf{K}_m$ (where the m refers to the number of modalities) which has the tableau system given by the rules in Figures B.3, B.4, and B.5 and for which the corresponding set of frames is K. Figure B.3 gives the basic tableau rules for handling the connectives, modalities, quantifiers, and abstractions. Figure B.4 gives the rules for equality. Figure B.5 gives the two rules for handling the global and local assumptions. Generally speaking, these rules are well known, but the versions here differ in some details, in particular, in the use of the admissibility assumption in the universal, abstraction, and substitutivity rules. The proof is completed when each branch is closed, that is, contains contradictory biterms.

Without loss of generality, all local and global assumptions are assumed to be closed, as is the theorem to be proved. Note, however, that during construction of a tableau, open biterms can be introduced by some tableau rules.

Definition B.3.4. Let $\mathcal{T}$ be a theory. A *proof of rank* 0 *with respect to* $\mathcal{T}$ is a sequence $T_1, \ldots, T_n$ of trees labelled by prefixed biterms satisfying the following conditions.

1. T_1 consists of a single node labelled by $1\ \neg\varphi$, for some biterm φ.
2. For $i = 1, \ldots, n-1$, there is
   (a) a tableau rule R from Figure B.3, Figure B.4, or Figure B.5 such that T_{i+1} is obtained from T_i,
      i. if R is a conjunctive rule, by extending a branch with two nodes labelled by the prefixed biterms in the denominator of R,
      ii. if R is a disjunctive rule, by splitting a branch so that the leaf node of the branch has two children each labelled by one of the prefixed biterms in the denominator of R,
      iii. otherwise, by extending a branch with a node labelled by the prefixed biterm in the denominator of R,

   provided that any prefixed biterms in the numerator of R already appear in the branch and any side-conditions of R are satisfied.
3. Each branch of T_n contains nodes labelled by $\sigma\ \psi$ and $\sigma\ \neg\psi$, for some prefix σ and biterm ψ.

Each T_i is called a *tableau of rank* 0.

A branch of a tableau of rank 0 is *closed* if it contains nodes labelled by $\sigma\ \psi$ and $\sigma\ \neg\psi$, for some prefix σ and biterm ψ; otherwise, the branch is *open*.

A tableau of rank 0 is *closed* if each branch is closed; otherwise, the tableau is *open*.

The biterm φ is called the *theorem* of the proof; this is denoted by $\mathcal{T} \vdash \varphi$.

The next task is to prove the soundness of the proof system. The key result that needs to be established is that, if a biterm has a proof using the rules in Figures B.3, B.4, and

(Conjunctive rules) For any prefix σ,

$$\frac{\sigma\ \ \varphi \wedge \psi}{\begin{array}{c}\sigma\ \ \varphi\\ \sigma\ \ \psi\end{array}} \qquad \frac{\sigma\ \ \neg(\varphi \vee \psi)}{\begin{array}{c}\sigma\ \ \neg\varphi\\ \sigma\ \ \neg\psi\end{array}} \qquad \frac{\sigma\ \ \neg(\varphi \longrightarrow \psi)}{\begin{array}{c}\sigma\ \ \varphi\\ \sigma\ \ \neg\psi\end{array}}$$

(Disjunctive rules) For any prefix σ,

$$\frac{\sigma\ \ \varphi \vee \psi}{\sigma\ \ \varphi \mid \sigma\ \ \psi} \qquad \frac{\sigma\ \ \neg(\varphi \wedge \psi)}{\sigma\ \ \neg\varphi \mid \sigma\ \ \neg\psi} \qquad \frac{\sigma\ \ \varphi \longrightarrow \psi}{\sigma\ \ \neg\varphi \mid \sigma\ \ \psi}$$

(Double negation rule) For any prefix σ,

$$\frac{\sigma\ \ \neg\neg\varphi}{\sigma\ \ \varphi}$$

(Possibility rules) If the prefix $\sigma.n_i$ is new to the branch, where $i \in \{1, \ldots, m\}$,

$$\frac{\sigma\ \ \Diamond_i\varphi}{\sigma.n_i\ \ \varphi} \qquad \frac{\sigma\ \ \neg\Box_i\varphi}{\sigma.n_i\ \ \neg\varphi}$$

(Necessity rules) If the prefix $\sigma.n_i$ already occurs on the branch, where $i \in \{1, \ldots, m\}$,

$$\frac{\sigma\ \ \Box_i\varphi}{\sigma.n_i\ \ \varphi} \qquad \frac{\sigma\ \ \neg\Diamond_i\varphi}{\sigma.n_i\ \ \neg\varphi}$$

(Σ rules) For any prefix σ, if $y_1, \ldots, y_n$ are variables new to the branch,

$$\frac{\sigma\ \ (\Sigma\ \varphi)}{\sigma\ \ \varphi\ y_1\ \ldots\ y_n} \qquad \frac{\sigma\ \ \neg(\Pi\ \varphi)}{\sigma\ \ \neg\varphi\ y_1\ \ldots\ y_n}$$

(Π rules) For any prefix σ, if $t_1, \ldots, t_k$ $(0 \leq k \leq n)$ are terms,

$$\frac{\sigma\ \ (\Pi\ \varphi)}{\sigma\ \ \varphi\ t_1\ \ldots\ t_k} \qquad \frac{\sigma\ \ \neg(\Sigma\ \varphi)}{\sigma\ \ \neg\varphi\ t_1\ \ldots\ t_k}$$

(Abstraction rules) For any prefix σ, if $\{x/t\}$ is admissible with respect to φ,

$$\frac{\sigma\ \ (\lambda x.\varphi\ \ t)}{\sigma\ \ \varphi\{x/t\}} \qquad \frac{\sigma\ \ \neg(\lambda x.\varphi\ \ t)}{\sigma\ \ \neg\varphi\{x/t\}}$$

Figure B.3: Basic tableau rules

(Reflexivity rule) If the prefix σ already occurs on the branch and t is a term,

$$\frac{}{\sigma \;\; t = t}$$

(Substitutivity rule) For any prefix σ, if φ is a biterm containing a free occurrence of the variable x, and $\{x/s\}$ and $\{x/t\}$ are admissible with respect to φ,

$$\frac{\sigma \;\; s = t \qquad \sigma \;\; \varphi\{x/s\}}{\sigma \;\; \varphi\{x/t\}}$$

Figure B.4: Tableau rules for equality

(Global assumption rule) If the prefix σ already occurs on the branch, ψ is a global assumption, and $t_1, \ldots, t_k$ $(0 \leq k \leq n)$ are terms,

$$\frac{}{\sigma \;\; \psi \; t_1 \; \ldots \; t_k}$$

(Local assumption rule) If ψ is a local assumption and $t_1, \ldots, t_k$ $(0 \leq k \leq n)$ are terms,

$$\frac{}{1 \;\; \psi \; t_1 \; \ldots \; t_k}$$

Figure B.5: Tableau rules for global and local assumptions

B.5, then the biterm is a K-consequence of the global and local assumptions. To this end, the concept of a satisfiable set of prefixed biterms is introduced.

Definition B.3.5. Let L be a class of frames. A set S of prefixed biterms is L-*satisfiable* with respect to a set $\mathcal{G}$ of global assumptions and a set $\mathcal{L}$ of local assumptions if there exists an interpretation $I \triangleq \langle W, \{R_i\}_{i=1}^m, \{\mathcal{D}_\alpha\}_{\alpha \in \mathfrak{S}}, \; V \rangle$ based on a frame in L, a variable assignment ν, and a mapping F of prefixes to worlds such that the following hold.

1. Each $\psi \in \mathcal{G}$ is valid in I.
2. Each $\psi \in \mathcal{L}$ is valid at $F(1)$ in I.
3. If the prefixes σ and $\sigma.n_i$ both occur in S, then $F(\sigma) \; R_i \; F(\sigma.n_i)$.
4. For every biterm type $\alpha \triangleq \alpha_1 \to \cdots \to \alpha_n \to o$, there exist $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$ such that if $\sigma \; \varphi$ is in S and φ has type α, then $\mathcal{V}(\varphi, I, F(\sigma), \nu) \; d_1 \; \ldots \; d_n = \mathsf{T}$.

A tableau branch is said to be L-*satisfiable* if the set of prefixed biterms on it is L-satisfiable. A tableau is said to be L-*satisfiable* if it has a branch that is L-satisfiable.

Notation. In the following, 'satisfiable' will mean 'K-satisfiable'.

Proposition B.3.3. *A closed tableau of rank 0 is not satisfiable.*

Proof. Suppose some closed tableau is satisfiable. Thus some branch of it must be satisfiable. Let S be the set of prefixed biterms on this branch. Suppose S is satisfiable using the interpretation I, variable assignment ν, and mapping F. Since the tableau is closed, the branch is closed and there exist prefixed biterms $\sigma\ \psi$ and $\sigma\ \neg\psi$ on the branch. Let the type of ψ be $\alpha_1 \to \cdots \to \alpha_n \to o$ and the domain elements satisfying Condition 4 of Definition B.3.5 for this type be $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$. Thus $\mathcal{V}(\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$ and $\mathcal{V}(\neg\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. However, by Part 4 of Proposition B.2.1, $\mathcal{V}(\neg\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$ iff $\mathcal{V}(\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{F}$, which gives a contradiction. Thus the closed tableau is not satisfiable. □

Since, for any class L of frames, an L-satisfiable set of prefixed biterms is satisfiable, it follows from Proposition B.3.3 that a closed tableau is not L-satisfiable.

Proposition B.3.4. *If a tableau rule (from Figure B.3, B.4, or B.5) is applied to a satisfiable tableau, then the resulting tableau is satisfiable.*

Proof. Suppose that $\mathcal{T}$ is a satisfiable tableau and some tableau rule is applied to it on branch $\mathcal{B}$. Suppose that $\mathcal{B}$ is not satisfiable. Thus some other branch of $\mathcal{T}$ is satisfiable and the tableau rule has no effect on this other branch. Hence the new tableau is satisfiable after the application of the tableau rule.

Thus it can be assumed that $\mathcal{B}$ is satisfiable. Suppose the set of prefixed biterms on $\mathcal{B}$ is satisfiable via the interpretation I, variable assignment ν, and mapping F. Also suppose the type of φ and ψ is $\alpha_1 \to \cdots \to \alpha_n \to o$ and the domain elements satisfying Condition 4 of Definition B.3.5 for this type are $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$. Each of the tableau rules is now examined in turn.

Conjunctive rules Consider the first of these. Since the original tableau is satisfiable, it follows that $\mathcal{V}(\varphi \wedge \psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Thus $\mathcal{V}(\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$ and $\mathcal{V}(\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$, by Part 1 of Proposition B.2.1. Thus the tableau extended by this rule is satisfiable.

The proofs of the other conjunctive rules are similar.

Disjunctive rules Consider the first of these. Since the original tableau is satisfiable, it follows that $\mathcal{V}(\varphi \vee \psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Thus $\mathcal{V}(\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$ or $\mathcal{V}(\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$, by Part 2 of Proposition B.2.1. Thus the tableau extended by this rule is satisfiable because at least one of the resulting branches is satisfiable.

The proofs of the other disjunctive rules are similar.

Double negation rule Since the original tableau is satisfiable, it follows that $\mathcal{V}(\neg\neg\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Then $\mathcal{V}(\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$, by Parts 5 and 6 of Proposition B.2.1. Thus the extended tableau is also satisfiable.

Possibility rules Consider the first of these. Since the original tableau is satisfiable, it follows that $\mathcal{V}(\Diamond_i\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. By Part 2 of Proposition B.2.6, for some w' such that $F(\sigma)\ R_i\ w'$, $\mathcal{V}(\varphi, I, w', \nu)\ d_1 \ldots d_n = \mathsf{T}$. Now extend the definition of F so that $F(\sigma.n_i) = w'$. Thus $F(\sigma)\ R_i\ F(\sigma.n_i)$ and $\mathcal{V}(\varphi, I, F(\sigma.n_i), \nu)\ d_1 \ldots d_n = \mathsf{T}$, and so the extended tableau is satisfiable.

The proof of the other possibility rule is similar.

Necessity rules

Consider the first of these. Since the original tableau is satisfiable, it follows that $\mathcal{V}(\Box_i\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Since the prefix $\sigma.n_i$ already occurs on the branch,

$F(\sigma)\ R_i\ F(\sigma.n_i)$, and so $\mathcal{V}(\varphi, I, F(\sigma.n_i), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Thus the extended tableau is satisfiable.

The proof of the other necessity rule is similar.

$\underline{\Sigma \text{ rules}}$ Consider the first of these. Since the original tableau is satisfiable, it follows that $\mathcal{V}((\Sigma\ \varphi), I, F(\sigma), \nu) = \mathsf{T}$. By Part 1 of Proposition B.2.2, $\mathcal{V}((\Sigma\ \varphi), I, F(\sigma), \nu) = \mathsf{T}$ iff, for some $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, $\mathcal{V}(\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Let y_i be a variable of type α_i that is new to the branch $(i = 1, \ldots, n)$. Define the variable assignment ν^* by $\nu^*(x) = \nu(x)$, for $x \notin \{y_1, \ldots, y_n\}$ and $\nu^*(y_i) = d_i$, for $i = 1, \ldots, n$. Since ν and ν^* agree on the free variables of φ, $\mathcal{V}(\varphi, I, F(\sigma), \nu) = \mathcal{V}(\varphi, I, F(\sigma), \nu^*)$, by Proposition B.2.14. Thus $\mathcal{V}(\varphi\ y_1 \ldots y_n, I, F(\sigma), \nu^*) = \mathcal{V}(\varphi, I, F(\sigma), \nu^*)\ d_1 \ldots d_n = \mathcal{V}(\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Now, since ν and ν^* differ only on the y_i and the y_i are new to the branch, ν and ν^* agree on the free variables of each biterm on the original branch. By Proposition B.2.14, it follows that the extended branch is satisfiable via the interpretation I, variable assignment ν^*, and mapping F.

The proof of the other Σ rule is similar, using Proposition B.2.3.

$\underline{\Pi \text{ rules}}$ Consider the first of these. Since the original tableau is satisfiable, it follows that $\mathcal{V}((\Pi\ \varphi), I, F(\sigma), \nu) = \mathsf{T}$. Hence, by Part 2 of Proposition B.2.2, $\mathcal{V}((\Pi\ \varphi), I, F(\sigma), \nu) = \mathsf{T}$ iff, for each $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, $\mathcal{V}(\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Let t_i be terms of type α_i and $\mathcal{V}(t_i, I, F(\sigma), \nu) = d_i$ $(i = 1, \ldots, k)$. Then we have that $\mathcal{V}(\varphi\ t_1 \ldots t_k, I, F(\sigma), \nu)\ d_{k+1} \ldots d_n = \mathcal{V}(\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$, for each $d_{k+1} \ldots d_n$. If this is the first occurrence on the branch of a biterm having the same type as $\varphi\ t_1 \ldots t_k$, then the $d_{k+1} \ldots d_n$ required by Condition 4 of Definition B.3.5 can be chosen arbitrarily; otherwise, the $d_{k+1} \ldots d_n$ have already been fixed and they work here as well. Hence the extended branch is satisfiable via the interpretation I, variable assignment ν, and mapping F.

The proof for the other Π rule is similar, using Proposition B.2.3.

$\underline{\text{Abstraction rules}}$ Consider the first of these. To show satisfiability of the tableau extended by this rule, it suffices to show that $\mathcal{V}(\varphi\{x/t\}, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Since the original tableau is satisfiable, it follows that $\mathcal{V}((\lambda x.\varphi\ t), I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. The result now follows immediately, since by Proposition B.2.33, $\mathcal{V}((\lambda x.\varphi\ t), I, F(\sigma), \nu) = \mathcal{V}(\varphi\{x/t\}, I, F(\sigma), \nu)$.

The proof for the second of the abstraction rules is similar.

$\underline{\text{Reflexivity rule}}$ To show satisfiability of the tableau extended by this rule, it suffices to show that $\mathcal{V}((t = t), I, F(\sigma), \nu) = \mathsf{T}$. But this follows immediately since the meaning of $=$ is the identity mapping.

$\underline{\text{Substitutivity rule}}$ To show satisfiability of the tableau extended by this rule, it suffices to show that $\mathcal{V}(\varphi\{x/t\}, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Since the original tableau is satisfiable, it follows that $\mathcal{V}(s = t, I, F(\sigma), \nu) = \mathsf{T}$ and $\mathcal{V}(\varphi\{x/s\}, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. The result now follows immediately from Proposition B.2.29.

$\underline{\text{Global assumption rule}}$ Since the (original) tableau is satisfiable and ψ is a global assumption, ψ is valid in I. Thus $\mathcal{V}(\psi, I, F(\sigma), \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$. Hence, for each $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, $\mathcal{V}(\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Let t_i be terms of type α_i and $\mathcal{V}(t_i, I, F(\sigma), \nu) = d_i$ $(i = 1, \ldots, k)$. Then $\mathcal{V}(\psi\ t_1 \ldots t_k, I, F(\sigma), \nu)\ d_{k+1} \ldots d_n = \mathcal{V}(\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \mathsf{T}$. If this is the first occurrence on the branch of a biterm having the same type as $\varphi\ t_1 \ldots t_k$, then the $d_{k+1} \ldots d_n$ required by Condition 4 of Definition B.3.5 can be chosen arbitrarily; otherwise, the $d_{k+1} \ldots d_n$ have already been fixed

and they work here as well. Hence the extended branch is satisfiable via the interpretation I, variable assignment ν, and mapping F.

Local assumption rule Since the (original) tableau is satisfiable and ψ is a local assumption, ψ is valid at $F(1)$ in I. Thus $\mathcal{V}(\psi, I, F(1), \nu) = \lambda x_1. \ldots \lambda x_n.\mathsf{T}$. Hence, for each $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$, $\mathcal{V}(\psi, I, F(1), \nu)\ d_1 \ldots d_n = \mathsf{T}$. Let t_i be terms of type α_i and $\mathcal{V}(t_i, I, F(1), \nu) = d_i$ $(i = 1, \ldots, k)$. Then $\mathcal{V}(\psi\ t_1 \ldots t_k, I, F(1), \nu)\ d_{k+1} \ldots d_n = \mathcal{V}(\psi, I, F(1), \nu)\ d_1 \ldots d_n = \mathsf{T}$. If this is the first occurrence on the branch of a biterm having the same type as $\varphi\ t_1 \ldots t_k$, then the $d_{k+1} \ldots d_n$ required by Condition 4 of Definition B.3.5 can be chosen arbitrarily; otherwise, the $d_{k+1} \ldots d_n$ have already been fixed and they work here as well. Hence the extended branch is satisfiable via the interpretation I, variable assignment ν, and mapping F.

□

Now the soundness result can be proved.

Proposition B.3.5. *Let $\mathcal{T}$ be a theory. Then the theorem of a proof of rank 0 with respect to $\mathcal{T}$ is a consequence of $\mathcal{T}$.*

Proof. Let $\mathcal{T}$ be $(\mathcal{G}, \mathcal{L})$. Suppose that φ is the theorem of a proof of rank 0 with respect to $\mathcal{T}$, but is not a consequence of $\mathcal{T}$. Since φ has a proof, there is a closed tableau T beginning with the prefixed biterm $1\ \neg\varphi$. Let T_1 be the tableau consisting of the single node $1\ \neg\varphi$. Thus T results from T_1 by applying various tableau rules in Figures B.3, B.4, and B.5.

Suppose that the type of φ is $\alpha_1 \to \cdots \to \alpha_n \to o$. Since φ is not a consequence, there is an interpretation I, a world w in I, a variable assignment ν, and $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$ such that each $\psi \in \mathcal{G}$ is valid in I, each $\psi \in \mathcal{L}$ is valid at w in I, and $\mathcal{V}(\varphi, I, w, \nu)\ d_1 \ldots d_n = \mathsf{F}$. Define the mapping F by $F(1) = w$. Clearly, $\{1\ \neg\varphi\}$ is satisfiable, using I, ν, and F, where d_i $(i = 1, \ldots, n)$ are the domain elements associated with the type of φ, as required by Condition 4 of Definition B.3.5. Thus T_1 is satisfiable.

Since T_1 is satisfiable, so is any tableau obtained from T_1 by applying tableau rules, by Proposition B.3.4. Thus T is satisfiable. But this gives a contradiction, by Proposition B.3.3, since T is closed. Thus φ is a consequence of $\mathcal{T}$. □

The next result is an immediate corollary of Proposition B.3.5.

Proposition B.3.6. *If a biterm φ has a proof of rank 0 with respect to the empty theory, then φ is valid.*

Figure B.6 gives the derived rules for the quantifiers that follow directly from rules given in Figure B.3. The most common case is given where φ is a formula. The rules can be formulated more generally, but the preference here is to keep the simpler versions instead and go back to the rules in Figure B.3 for more general situations.

Sometimes it will be convenient to have a further rule available that introduces $\lambda x_1. \ldots \lambda x_n.\top$ or $\neg\lambda x_1. \ldots \lambda x_n.\bot$ $(n \geq 0)$ into a proof, as in Figure B.7. The proof of correctness of this rule is almost identical to that of the reflexivity rule. It enables a branch to be closed as soon as either of the prefixed biterms $\sigma\ \ \lambda x_1. \ldots \lambda x_n.\bot$ or $\sigma\ \ \neg\lambda x_1. \ldots \lambda x_n.\top$ appear in it.

To keep proofs as compact as possible, it is convenient to have some derived rules. The next result shows the correctness of the derived rules in Figure B.8 that are useful since assumptions often have the implicational form $\varphi \longrightarrow \psi$.

(Existential rules) For any prefix σ, if φ is a formula and y is a variable new to the branch,

$$\frac{\sigma \;\; \exists x.\varphi}{\sigma \;\; \varphi\{x/y\}} \qquad \frac{\sigma \;\; \neg\forall x.\varphi}{\sigma \;\; \neg\varphi\{x/y\}}$$

(Universal rules) For any prefix σ, if $\{x/t\}$ is admissible with respect to the formula φ,

$$\frac{\sigma \;\; \forall x.\varphi}{\sigma \;\; \varphi\{x/t\}} \qquad \frac{\sigma \;\; \neg\exists x.\varphi}{\sigma \;\; \neg\varphi\{x/t\}}$$

Figure B.6: Derived rules for existential and universal quantifiers

($\top$ introduction rules) For any prefix σ,

$$\frac{}{\sigma \;\; \lambda x_1.\ldots.\lambda x_n.\top} \qquad \frac{}{\sigma \;\; \neg\lambda x_1.\ldots.\lambda x_n.\bot}$$

Figure B.7: $\lambda x_1.\ldots.\lambda x_n.\top$ introduction rules

Proposition B.3.7.

1. *Use of the global assumption $\varphi \longrightarrow \psi$ is equivalent to use of either of the tableau rules:*

$$\frac{\sigma \;\; \varphi}{\sigma \;\; \psi} \qquad \frac{\sigma \;\; \neg\psi}{\sigma \;\; \neg\varphi}$$

2. *Use of the local assumption $\varphi \longrightarrow \psi$ is equivalent to use of either of the tableau rules:*

$$\frac{1 \;\; \varphi}{1 \;\; \psi} \qquad \frac{1 \;\; \neg\psi}{1 \;\; \neg\varphi}$$

Proof. 1. Suppose first that the tableau rule $\dfrac{\sigma \;\; \varphi}{\sigma \;\; \psi}$ is available. It is shown that this makes $\varphi \longrightarrow \psi$ available as a global assumption. Suppose that σ is a prefix already occurring on some branch. Consider the prefixed biterm

$$\sigma \;\; (\varphi \longrightarrow \psi) \vee \neg(\varphi \longrightarrow \psi).$$

Since the biterm is equivalent to $\lambda x_1.\ldots.\lambda x_n.\top$, this could be added (correctly) to the end of the branch. A disjunctive rule can then be used to produce two children

$$\sigma \;\; \varphi \longrightarrow \psi \quad \text{and} \quad \sigma \;\; \neg(\varphi \longrightarrow \psi).$$

From the second of these, the tableau proof can proceed as follows.

$\sigma \;\; \neg(\varphi \longrightarrow \psi) \;\; 1.$
$\sigma \;\; \varphi \;\; 2.$
$\sigma \;\; \neg\psi \;\; 3.$
$\sigma \;\; \psi \;\; 4.$

Items 2 and 3 are from 1 by a conjunctive rule; 4 is from 2 by the proposed rule. Now this branch closes by 3 and 4. What remains in the tableau is the prefixed biterm $\sigma\ \ \varphi \longrightarrow \psi$, which is exactly what would result from the use of $\varphi \longrightarrow \psi$ as a global assumption.

For the converse, suppose that $\varphi \longrightarrow \psi$ is a global assumption. It has to be shown that the use of the proposed rule $\dfrac{\sigma\ \ \varphi}{\sigma\ \ \psi}$ is sound, that is, if the rule is applied to a satisfiable tableau, then the resulting tableau is satisfiable. Suppose the rule is applied to a branch $\mathcal{B}$. As in the proof of Proposition B.3.4, it can be assumed that $\mathcal{B}$ is satisfiable. Suppose the set of prefixed biterms on $\mathcal{B}$ is satisfiable via the interpretation I, variable assignment ν, and mapping F. Since the (original) tableau is satisfiable, there exists $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$ such that $\mathcal{V}(\varphi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \top$. Also $\mathcal{V}(\varphi \longrightarrow \psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \top$, since $\varphi \longrightarrow \psi$ is a global assumption. Thus $\mathcal{V}(\psi, I, F(\sigma), \nu)\ d_1 \ldots d_n = \top$, by Part 3 of Proposition B.2.1, as required.

The proof for the rule $\dfrac{\sigma\ \ \neg\psi}{\sigma\ \ \neg\varphi}$ is similar.

2. The proof is similar to that of 1. □

(Derived rule for global implicational assumption) For any prefix σ, if $\varphi \longrightarrow \psi$ is a global assumption,

$$\frac{\sigma\ \ \varphi}{\sigma\ \ \psi} \qquad\qquad \frac{\sigma\ \ \neg\psi}{\sigma\ \ \neg\varphi}$$

(Derived rule for local implicational assumption) If $\varphi \longrightarrow \psi$ is a local assumption,

$$\frac{1\ \ \varphi}{1\ \ \psi} \qquad\qquad \frac{1\ \ \neg\psi}{1\ \ \neg\varphi}$$

Figure B.8: Derived rules for implicational assumptions

Note that, in the case of theorem proving, the definition of *if_then_else* given in standard equality theory can be replaced by the following more general theory.

$$x \longrightarrow \textit{if } x \textit{ then } y \textit{ else } z = y$$
$$\neg x \longrightarrow \textit{if } x \textit{ then } y \textit{ else } z = z.$$

Not surprisingly, the result of a computation can also be obtained by proving a theorem. This is shown by simulating a computation step using a proof.

Proposition B.3.8. *Let $\mathcal{T} \triangleq (\mathcal{G}, \mathcal{L})$ be a theory, t a term, and s a subterm of t at occurrence o, where the modal path to o in t is $k_1 \ldots k_m$. Suppose there is*

1. *a formula $\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_m} \forall(u = v)$ in $\mathcal{L}$, or*
2. *a formula $\forall(u = v)$ in $\mathcal{G}$, and*

a substitution θ that is admissible with respect to $u = v$ such that $u\theta$ is α-equivalent to s. Then $\mathcal{T} \vdash \Box_{j_1} \cdots \Box_{j_r} \forall(t = t[s/v\theta]_o)$.

Proof. First, the case when s is t is considered. Thus o is ε, the modal path is empty, and $u\theta$ is α-equivalent to t. It has to be shown that $\mathcal{T} \vdash \Box_{j_1} \cdots \Box_{j_r} \forall (u\theta = v\theta)$. The proof of this is given in Figure B.9. An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is a local assumption; 3 is from 1 by possibility rules; 4 is from 3 by existential rules (where φ is the substitution introducing the variables new to the branch); 5 is from 2 by necessity rules; 6 is from 5 by universal rules; now the branch closes by 4 and 6. The proof when the input equation is a global assumption is almost the same as this.

1	$\neg\Box_{j_1} \cdots \Box_{j_r} \forall (u\theta = v\theta)$ 1.
1	$\Box_{j_1} \cdots \Box_{j_r} \forall (u = v)$ 2.
$1.1_{j_1} \ldots 1_{j_r}$	$\neg\forall (u\theta = v\theta)$ 3.
$1.1_{j_1} \ldots 1_{j_r}$	$\neg(u\theta\varphi = v\theta\varphi)$ 4.
$1.1_{j_1} \ldots 1_{j_r}$	$\forall (u = v)$ 5.
$1.1_{j_1} \ldots 1_{j_r}$	$u\theta\varphi = v\theta\varphi$ 6.

Figure B.9: Proof of $\Box_{j_1} \cdots \Box_{j_r} \forall (u\theta = v\theta)$

Now the case when s is a *strict* subterm of t is considered. The proof for this case is by induction on the structure of t. If t is a variable or a constant, then there are no strict subterms and there is nothing more to consider.

Suppose that t is an abstraction $\lambda x.r$. Since s is a strict subterm of t, s must be a subterm of r at occurrence o', where $o = 1o'$. By either the induction hypothesis or the first part of the proof in case s is r, we have that $\mathcal{T} \vdash \Box_{j_1} \cdots \Box_{j_r} \forall (r = r[s/v\theta]_{o'})$. It has to be shown that $\mathcal{T} \vdash \Box_{j_1} \cdots \Box_{j_r} \forall (\lambda x.r = (\lambda x.r)[s/v\theta]_o)$. The proof of this is given in Figure B.10. An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is from the induction hypothesis; 3 is from 1 by possibility rules; 4 is from 3 by existential rules (where φ is the substitution introducing the variables new to the branch); 5 is a global assumption (the axiom of extensionality); 6 is from 5 by universal rules; 7 is from 4 and 6 by the substitutivity rule; 8 is from 7 by an existential rule (z is a variable new to the branch); 9 is a global assumption (β-reduction); 10 is a global assumption (β-reduction); 11 is from 8 and 9 by the substitutivity rule; 12 is from 10 and 11 by the substitutivity rule; 13 is from 2 by necessity rules; 14 is from 13 by universal rules; now the branch closes by 12 and 14.

Suppose that t is a modal term $(t_1\ t_2)$. Since s is a strict subterm of t, s either must be a subterm of t_1 at occurrence o', where $o = 1o'$, or a subterm of t_2 at occurrence o', where $o = 2o'$. The cases are essentially the same, so just the first is considered. By either the induction hypothesis or the first part of the proof in case s is t_1, we have that $\mathcal{T} \vdash \Box_{j_1} \cdots \Box_{j_r} \forall (t_1 = t_1[s/v\theta]_{o'})$. It has to be shown that $\mathcal{T} \vdash \Box_{j_1} \cdots \Box_{j_r} \forall ((t_1\ t_2) = (t_1\ t_2)[s/v\theta]_o)$. The proof of this is given in Figure B.11. An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is from the induction hypothesis; 3 is from 1 by possibility rules; 4 is from 3 by existential rules (where φ is the substitution introducing the variables new to the branch); 5 is from 2 by necessity rules; 6 is from 5 by universal rules; 7 is by the reflexivity rule; 8 is from 6 and 7 by the substitutivity rule; now the branch closes by 4 and 8.

Suppose that t is a modal term $(t_1, \ldots, t_n)$. The proof for this case is similar to the

1	$\neg\Box_{j_1}\cdots\Box_{j_r}\forall(\lambda x.r = (\lambda x.r)[s/v\theta]_o)$ 1.
1	$\Box_{j_1}\cdots\Box_{j_r}\forall(r = r[s/v\theta]_{o'})$ 2.
$1.1_{j_1}\ldots 1_{j_r}$	$\neg\forall(\lambda x.r = (\lambda x.r)[s/v\theta]_o)$ 3.
$1.1_{j_1}\ldots 1_{j_r}$	$\neg((\lambda x.r)\varphi = ((\lambda x.r)[s/v\theta]_o)\varphi)$ 4.
$1.1_{j_1}\ldots 1_{j_r}$	$\forall f.\forall g.((f = g) = \forall y.((f\ y) = (g\ y)))$ 5.
$1.1_{j_1}\ldots 1_{j_r}$	$((\lambda x.r)\varphi = ((\lambda x.r)[s/v\theta]_o)\varphi) =$ $\forall y.(((\lambda x.r)\varphi\ y) = (((\lambda x.r)[s/v\theta]_o)\varphi\ y))$ 6.
$1.1_{j_1}\ldots 1_{j_r}$	$\neg\forall y.(((\lambda x.r)\varphi\ y) = (((\lambda x.r)[s/v\theta]_o)\varphi\ y))$ 7.
$1.1_{j_1}\ldots 1_{j_r}$	$\neg(((\lambda x.r)\varphi\ z) = (((\lambda x.r)[s/v\theta]_o)\varphi\ z))$ 8.
$1.1_{j_1}\ldots 1_{j_r}$	$((\lambda x.r)\varphi\ z) = r\varphi\{x/z\}$ 9.
$1.1_{j_1}\ldots 1_{j_r}$	$(((\lambda x.r)[s/v\theta]_o)\varphi\ z) = r[s/v\theta]_{o'}\varphi\{x/z\}$ 10.
$1.1_{j_1}\ldots 1_{j_r}$	$\neg(r\varphi\{x/z\} = (((\lambda x.r)[s/v\theta]_o)\varphi\ z))$ 11.
$1.1_{j_1}\ldots 1_{j_r}$	$\neg(r\varphi\{x/z\} = r[s/v\theta]_{o'}\varphi\{x/z\})$ 12.
$1.1_{j_1}\ldots 1_{j_r}$	$\forall(r = r[s/v\theta]_{o'})$ 13.
$1.1_{j_1}\ldots 1_{j_r}$	$r\varphi\{x/z\} = r[s/v\theta]_{o'}\varphi\{x/z\}$ 14.

Figure B.10: Proof of $\Box_{j_1}\cdots\Box_{j_r}\forall(\lambda x.r = (\lambda x.r)[s/v\theta]_o)$

1	$\neg\Box_{j_1}\cdots\Box_{j_r}\forall((t_1\ t_2) = (t_1\ t_2)[s/v\theta]_o)$ 1.
1	$\Box_{j_1}\cdots\Box_{j_r}\forall(t_1 = t_1[s/v\theta]_{o'})$ 2.
$1.1_{j_1}\ldots 1_{j_r}$	$\neg\forall((t_1\ t_2) = (t_1\ t_2)[s/v\theta]_o)$ 3.
$1.1_{j_1}\ldots 1_{j_r}$	$\neg((t_1\ t_2)\varphi = (t_1\ t_2)[s/v\theta]_o\varphi)$ 4.
$1.1_{j_1}\ldots 1_{j_r}$	$\forall(t_1 = t_1[s/v\theta]_{o'})$ 5.
$1.1_{j_1}\ldots 1_{j_r}$	$t_1\varphi = t_1[s/v\theta]_{o'}\varphi$ 6.
$1.1_{j_1}\ldots 1_{j_r}$	$(t_1\ t_2)\varphi = (t_1\ t_2)\varphi$ 7.
$1.1_{j_1}\ldots 1_{j_r}$	$(t_1\ t_2)\varphi = (t_1\ t_2)[s/v\theta]_o\varphi$ 8.

Figure B.11: Proof of $\Box_{j_1}\cdots\Box_{j_r}\forall((t_1\ t_2) = (t_1\ t_2)[s/v\theta]_o)$

preceding one.

Suppose that t is a modal term $\Box_{k_1}r$. Since s is a strict subterm of t, s must be a subterm of r at occurrence o', where $o = 1o'$. By either the induction hypothesis or the first part of the proof in case s is r, we have that $\mathfrak{T} \vdash \Box_{j_1}\cdots\Box_{j_r}\Box_{k_1}\forall(r = r[s/v\theta]_{o'})$. It has to be shown that $\mathfrak{T} \vdash \Box_{j_1}\cdots\Box_{j_r}\forall(\Box_{k_1}r = (\Box_{k_1}r)[s/v\theta]_o)$. The proof of this is given in Figure B.12. An explanation of this proof is as follows. Item 1 is the negation of the formula to be proved; 2 is from the induction hypothesis; 3 is from 1 by possibility rules; 4 is from 3 by existential rules (where φ is the substitution introducing the variables new to the branch); 5 is the global assumption $\Box_i(\boldsymbol{s} = \boldsymbol{t}) \longrightarrow (\Box_i\boldsymbol{s} = \Box_i\boldsymbol{t})$; 6 and 10 are from 5 by a disjunctive rule; 7 is from 6 by a possibility rule; 8 is from 2 by necessity rules; 9 is from 8 by a universal rule; now the first branch closes by 7 and 9, and the second closes by 4 and 10.

□

An easy induction argument using Proposition B.3.8 shows that the result of any computation is a theorem that can be proved without resorting to computation at all.

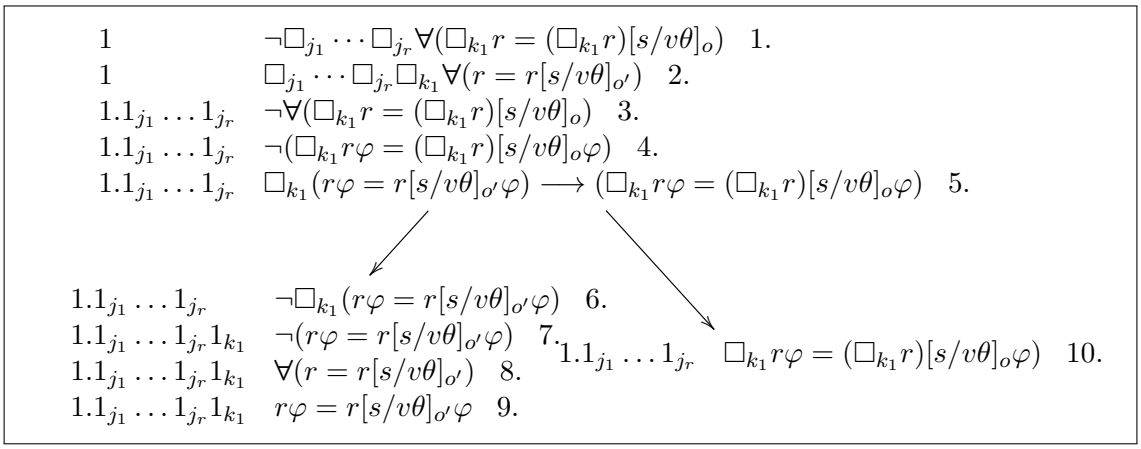

Figure B.12: Proof of $\Box_{j_1}\cdots\Box_{j_r}\forall(\Box_{k_1}r = (\Box_{k_1}r)[s/v\theta]_o)$

In other words, from a theoretical point of view, computation is redundant. However, in practice, computation is essential, for at least two related reasons. The first is that in typical applications the most common reasoning task is a computational one, that of evaluating a function on some argument. The second is that a proof that simulates a computation is generally considerably more complicated than the computation.

This section concludes with the tableau rules for the logic $\mathbf{KD45}_m$ that are given in Figure B.13.

($\mathbf{D}_i$ rule) If the prefix σ already occurs on the branch,

$$\frac{\sigma\ \ \Box_i\varphi}{\sigma\ \ \Diamond_i\varphi} \qquad \frac{\sigma\ \ \neg\Diamond_i\varphi}{\sigma\ \ \neg\Box_i\varphi}$$

($\mathbf{4}_i$ rule) If the prefixes σ and $\sigma.n_i$ already occur on the branch,

$$\frac{\sigma\ \ \Box_i\varphi}{\sigma.n_i\ \ \Box_i\varphi} \qquad \frac{\sigma\ \ \neg\Diamond_i\varphi}{\sigma.n_i\ \ \neg\Diamond_i\varphi}$$

($\mathbf{5}_i$ rule) If the prefixes σ and $\sigma.n_i$ already occur on the branch,

$$\frac{\sigma.n_i\ \ \Box_i\varphi}{\sigma\ \ \Box_i\varphi} \qquad \frac{\sigma.n_i\ \ \neg\Diamond_i\varphi}{\sigma\ \ \neg\Diamond_i\varphi}$$

Figure B.13: Tableau rules for the logic $\mathbf{KD45}_m$

Proposition B.3.9. *Suppose that the following are global assumptions of the theory:*

$\Box_i\varphi \longrightarrow \Diamond_i\varphi \quad (D_i)$
$\Box_i\varphi \longrightarrow \Box_i\Box_i\varphi \quad (4_i)$
$\neg\Box_i\varphi \longrightarrow \Box_i\neg\Box_i\varphi \quad (5_i)$

If a tableau rule from Figure B.13 is applied to a satisfiable tableau, then the resulting tableau is satisfiable.

Proof. Suppose that $\mathcal{T}$ is a satisfiable tableau and some tableau rule from Figure B.13 is applied to it on branch $\mathcal{B}$. Suppose that $\mathcal{B}$ is not satisfiable. Thus some other branch of $\mathcal{T}$ is satisfiable and the tableau rule has no effect on this other branch. Hence the new tableau is satisfiable after the application of the tableau rule.

Thus it can be assumed that $\mathcal{B}$ is satisfiable. Suppose the set of prefixed biterms on $\mathcal{B}$ is satisfiable via the interpretation I, variable assignment ν, and mapping F. Also suppose the type of φ is $\alpha_1 \to \cdots \to \alpha_n \to o$ and the domain elements satisfying Condition 4 of Definition B.3.5 for this type are $d_i \in \mathcal{D}_{\alpha_i}$ $(i = 1, \ldots, n)$. Each of the tableau rules in Figure B.13 is now examined in turn.

Consider first the $\mathbf{D}_i$ rule. Since the original tableau is satisfiable, it follows that

$$\mathcal{V}(\Box_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Also, since $\Box_i\varphi \longrightarrow \Diamond_i\varphi$ is a global assumption,

$$\mathcal{V}(\Box_i\varphi \longrightarrow \Diamond_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Hence

$$\mathcal{V}(\Diamond_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Thus the tableau extended by rule $\mathbf{D}_i$ is satisfiable.

Consider next the $\mathbf{4}_i$ rule. Since the original tableau is satisfiable, it follows that

$$\mathcal{V}(\Box_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Also, since $\Box_i\varphi \longrightarrow \Box_i\Box_i\varphi$ is a global assumption,

$$\mathcal{V}(\Box_i\varphi \longrightarrow \Box_i\Box_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Hence

$$\mathcal{V}(\Box_i\Box_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Since the prefix $\sigma.n_i$ already occurs on the branch, $F(\sigma)\ R_i\ F(\sigma.n_i)$ and so

$$\mathcal{V}(\Box_i\varphi, I, F(\sigma.n_i), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Thus the tableau extended by rule $\mathbf{4}_i$ is satisfiable.

Finally, consider the $\mathbf{5}_i$ rule. Since the original tableau is satisfiable, it follows that

$$\mathcal{V}(\Box_i\varphi, I, F(\sigma.n_i), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}$$

and hence

$$\mathcal{V}(\Diamond_i\Box_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Also, since $\Diamond_i\Box_i\varphi \longrightarrow \Box_i\varphi$ is a global assumption,

$$\mathcal{V}(\Diamond_i\Box_i\varphi \longrightarrow \Box_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T},$$

and so

$$\mathcal{V}(\Box_i\varphi, I, F(\sigma), \nu)\ d_1\ \ldots\ d_n = \mathsf{T}.$$

Thus the tableau extended by rule $\mathbf{5}_i$ is satisfiable. □

B.3.3 Computation and Proof Combined

The next step is to combine computation and proof to give the full reasoning system. Before that, some differences between computation and proof are pointed out.

The primary difference is that computation is concerned with computing the meaning of arbitrary terms, whereas proof is concerned with finding the meaning of biterms. There is also a difference in the kind of theories that each requires: computation needs (primarily) an equational theory, whereas proof works with arbitrary theories. Typically, computation is 'efficient'; indeed, it is the basis of declarative programming languages. Proof usually involves much more search and therefore is likely to be much less efficient.

This section defines the combination of proof and computation, and shows the usefulness of this combination. Computation enhances proof with a powerful equational reasoning system; proof enhances computation by allowing some of the theory to be not in equational form. Here now are the details. By means of two mutually recursive definitions, the concepts of computation of rank k and proof of rank k are defined, for $k \geq 1$.

Definition B.3.6. Let $\mathcal{T} \triangleq (\mathcal{G}, \mathcal{L})$ be a theory and $k \geq 1$. A *computation of rank k using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{T}$* is a sequence $\{t_i\}_{i=1}^{n}$ of terms such that, for $i = 1, \ldots, n-1$, there is

1. a subterm s_i of t_i at occurrence o_i, where the modal path to o_i in t_i is $k_1 \ldots k_{m_i}$,
2. (a) a formula $\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_i}} \forall (u_i = v_i)$ in $\mathcal{L}$, or
   (b) a formula $\forall (u_i = v_i)$ in $\mathcal{G}$, or
   (c) a formula $\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_i}} \forall (u_i = v_i)$ that is the result of a computation of rank $k-1$ using $\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_i}}$ with respect to $\mathcal{T}$, or
   (d) a formula $\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_i}} \forall (u_i = v_i)$ that is the theorem of a proof of rank $k-1$ with respect to $\mathcal{T}$, and
3. a substitution θ_i that is admissible with respect to $u_i = v_i$

such that $u_i\theta_i$ is α-equivalent to s_i and t_{i+1} is $t_i[s_i / v_i\theta_i]_{o_i}$.

The term t_1 is called the *goal* of the computation and t_n is called the *answer*.

Each subterm s_i is called a *redex*.

Each formula $\Box_{j_1} \cdots \Box_{j_r} \Box_{k_1} \cdots \Box_{k_{m_i}} \forall (u_i = v_i)$ or $\forall (u_i = v_i)$ in Part 2 of the definition is called an *input equation*.

The formula $\Box_{j_1} \cdots \Box_{j_r} \forall (t_1 = t_n)$ is called the *result* of the computation.

Definition B.3.7. Let $\mathcal{T} \triangleq (\mathcal{G}, \mathcal{L})$ be a theory and $k \geq 1$. A *proof of rank k with respect to $\mathcal{T}$* is a sequence $T_1, \ldots, T_n$ of trees labelled by prefixed biterms satisfying the following conditions.

1. T_1 consists of a single node labelled by $1\ \neg\varphi$, for some biterm φ.
2. For $i = 1, \ldots, n-1$, there is either
   (a) a tableau rule R from Figure B.3, Figure B.4, or Figure B.5 such that T_{i+1} is obtained from T_i,

i. if R is a conjunctive rule, by extending a branch with two nodes labelled by the prefixed biterms in the denominator of R,

ii. if R is a disjunctive rule, by splitting a branch so that the leaf node of the branch has two children each labelled by one of the prefixed biterms in the denominator of R,

iii. otherwise, by extending a branch with a node labelled by the prefixed biterm in the denominator of R,

provided that any prefixed biterms in the numerator of R already appear in the branch and any side-conditions of R are satisfied; or

(b) there is a theorem η of a proof of rank $k-1$ and a branch is extended with the prefixed biterm $1\ \eta$; or

(c) there is a result η of a computation of rank $k-1$ and a branch is extended with the prefixed biterm $1\ \eta$.

3. Each branch of T_n contains nodes labelled by $\sigma\ \psi$ and $\sigma\ \neg\psi$, for some prefix σ and biterm ψ.

Each T_i is called a *tableau of rank k*.

A branch of a tableau of rank k is *closed* if it contains nodes labelled by $\sigma\ \psi$ and $\sigma\ \neg\psi$, for some prefix σ and biterm ψ; otherwise, the branch is *open*.

A tableau of rank k is *closed* if each branch is closed; otherwise, the tableau is *open*.

The biterm φ is called the *theorem* of the proof; this is denoted by $\mathcal{T} \vdash \varphi$.

Note that a computation of rank k is a computation of rank k' and a proof of rank k is a proof of rank k', for all $k' > k$.

Definition B.3.8. Let $\mathcal{T}$ be a theory.

A *computation with respect to* $\mathcal{T}$ is a computation using $\Box_{j_1} \cdots \Box_{j_r}$ of rank k with respect to $\mathcal{T}$, for some $j_1 \ldots j_r$ and $k \geq 0$.

A *proof with respect to* $\mathcal{T}$ is a proof of rank k with respect to $\mathcal{T}$, for some $k \geq 0$.

Proposition B.3.10. *Let $\mathcal{T}$ be a theory. Then the following hold.*

1. *The result of a computation with respect to $\mathcal{T}$ is a consequence of $\mathcal{T}$.*

2. *The theorem of a proof with respect to $\mathcal{T}$ is a consequence of $\mathcal{T}$.*

Proof. The proof is by induction on the rank of the computation and the rank of the proof.

If the rank of the computation is 0, then this is just Proposition B.3.1, and if the rank of the proof is 0, then this is just Proposition B.3.5.

Suppose next that the result holds for computations of rank $k-1$ and proofs of rank $k-1$. Consider first a computation of rank k. By the induction hypothesis, the input equation in Parts 2(c) and 2(d) of the definition of computation of rank k is consequence of $\mathcal{T}$. Using this fact, the proof that a computation of rank k is a consequence of $\mathcal{T}$ is now almost exactly the same as the proof of Proposition B.3.1.

Finally, consider a proof of rank k. The first step is to establish that Proposition B.3.4 can be extended to the Parts 2(b) and 2(c) in the definition of proof of rank k in which a branch is extended with a prefixed biterm $1\ \eta$, where η is either theorem of a proof of rank

$k-1$ or a result of a computation of rank $k-1$. In either case, by the induction hypothesis, η is a consequence of $\mathcal{T}$, and so η is valid at $F(1)$ in I (where I is the interpretation in the proof of Proposition B.3.4). Thus the tableau extended by a node labelled by 1 η is satisfiable, which establishes Proposition B.3.4 for these two extra cases. The remainder of the proof that a proof of rank k is a consequence of $\mathcal{T}$ now follows in a similar way to the proof of Proposition B.3.5. □

In some applications, it is common for a proof to consist largely of a computation.

Proposition B.3.11. *Let $\mathcal{T}$ be a theory, φ a biterm of type $\alpha_1 \to \cdots \to \alpha_n \to o$, and $k \geq 0$.*

1. *Suppose there is a computation of rank k using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{T}$ of φ with result $\Box_{j_1} \cdots \Box_{j_r} \forall(\varphi = \lambda x_1 \ldots . \lambda x_n . \top)$. Then there is a proof of rank $k+1$ with respect to $\mathcal{T}$ of $\Box_{j_1} \cdots \Box_{j_r} \forall(\varphi)$.*

2. *Suppose there is a computation of rank k using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{T}$ of φ with result $\Box_{j_1} \cdots \Box_{j_r} \forall(\varphi = \lambda x_1 \ldots . \lambda x_n . \bot)$. Then there is a proof of rank $k+1$ with respect to $\mathcal{T}$ of $\Box_{j_1} \cdots \Box_{j_r} \forall(\neg\varphi)$.*

Proof. 1. The proof of $\Box_{j_1} \cdots \Box_{j_r} \forall(\varphi)$ is given in Figure B.14. Suppose that φ contains the free variables $y_1, \ldots, y_m$ ($m \geq 0$). Then φ' denotes the biterm $\varphi\{y_1/z_1, \ldots, y_m/z_m\}$, where $z_1, \ldots, z_m$ are variables new to the branch of the proof. An explanation of this proof is as follows. Item 1 is the negation of the biterm to be proved; 2 is from 1 by a possibility rule; 3 is from 2 by possibility rules; 4 is from 3 by existential rules; 5 is a result from a computation of rank k; 6 is from 5 by a necessity rule; 7 is from 6 by necessity rules; 8 is from 7 by universal rules; 9 is from 4 and 8 by the substitutivity rule; now the branch closes by 9.

2. The proof of $\Box_{j_1} \cdots \Box_{j_r} \forall(\neg\varphi)$ is given in Figure B.15. An explanation of this proof is as follows. Item 1 is the negation of the biterm to be proved; 2 is from 1 by a possibility rule; 3 is from 2 by possibility rules; 4 is from 3 by existential rules; 5 is from 4 by the double negation rule; 6 is a result from a computation of rank k; 7 is from 6 by a necessity rule; 8 is from 7 by necessity rules; 9 is from 8 by universal rules; 10 is from 5 and 9 by the substitutivity rule; now the branch closes by 10. □

The next result, which builds on Proposition B.3.2, is used in the theoretical development of belief acquisition.

Proposition B.3.12. *Let s and t be basic terms of the same type, $\mathcal{E}$ the standard equality theory, and $\Box_{j_1} \cdots \Box_{j_r}$ any sequence of modalities. Then the following hold.*

1. *If $s \overset{*}{\longleftrightarrow}_\alpha t$, then $\mathcal{E} \vdash \Box_{j_1} \cdots \Box_{j_r}(s = t)$.*

2. *If $s \not\overset{*}{\longleftrightarrow}_\alpha t$, then $\mathcal{E} \vdash \Box_{j_1} \cdots \Box_{j_r} \neg(s = t)$.*

Proof. 1. Suppose that $s \overset{*}{\longleftrightarrow}_\alpha t$. According to Proposition B.3.2, there is a computation of rank 0 using $\Box_{j_1} \cdots \Box_{j_r}$ with respect to $\mathcal{E}$ for the goal $s = t$ that has the result $\Box_{j_1} \cdots \Box_{j_r}((s = t) = \top)$. Using this, the proof of $\Box_{j_1} \cdots \Box_{j_r}(s = t)$ is given in Figure B.16. An explanation of this proof is as follows. Item 1 is the negation of the formula to be

1	$\neg\Box_{j_1}\cdots\Box_{j_r}\forall(\varphi)$	1.
1.1_{j_1}	$\neg\Box_{j_2}\cdots\Box_{j_r}\forall(\varphi)$	2.
	$\vdots$	
$1.1_{j_1}\ldots1_{j_r}$	$\neg\forall(\varphi)$	3.
$1.1_{j_1}\ldots1_{j_r}$	$\neg\varphi'$	4.
1	$\Box_{j_1}\cdots\Box_{j_r}\forall(\varphi=\lambda x_1.\ldots.\lambda x_n.\top)$	5.
1.1_{j_1}	$\Box_{j_2}\cdots\Box_{j_r}\forall(\varphi=\lambda x_1.\ldots.\lambda x_n.\top)$	6.
	$\vdots$	
$1.1_{j_1}\ldots1_{j_r}$	$\forall(\varphi=\lambda x_1.\ldots.\lambda x_n.\top)$	7.
$1.1_{j_1}\ldots1_{j_r}$	$\varphi'=\lambda x_1.\ldots.\lambda x_n.\top$	8.
$1.1_{j_1}\ldots1_{j_r}$	$\neg\lambda x_1.\ldots.\lambda x_n.\top$	9.

Figure B.14: Proof of rank $k+1$ of $\Box_{j_1}\cdots\Box_{j_r}\forall(\varphi)$

1	$\neg\Box_{j_1}\cdots\Box_{j_r}\forall(\neg\varphi)$	1.
1.1_{j_1}	$\neg\Box_{j_2}\cdots\Box_{j_r}\forall(\neg\varphi)$	2.
	$\vdots$	
$1.1_{j_1}\ldots1_{j_r}$	$\neg\forall(\neg\varphi)$	3.
$1.1_{j_1}\ldots1_{j_r}$	$\neg\neg\varphi'$	4.
$1.1_{j_1}\ldots1_{j_r}$	φ'	5.
1	$\Box_{j_1}\cdots\Box_{j_r}\forall(\varphi=\lambda x_1.\ldots.\lambda x_n.\bot)$	6.
1.1_{j_1}	$\Box_{j_2}\cdots\Box_{j_r}\forall(\varphi=\lambda x_1.\ldots.\lambda x_n.\bot)$	7.
	$\vdots$	
$1.1_{j_1}\ldots1_{j_r}$	$\forall(\varphi=\lambda x_1.\ldots.\lambda x_n.\bot)$	8.
$1.1_{j_1}\ldots1_{j_r}$	$\varphi'=\lambda x_1.\ldots.\lambda x_n.\bot$	9.
$1.1_{j_1}\ldots1_{j_r}$	$\lambda x_1.\ldots.\lambda x_n.\bot$	10.

Figure B.15: Proof of rank $k+1$ of $\Box_{j_1}\cdots\Box_{j_r}\forall(\neg\varphi)$

proved; 2 is the result from Proposition B.3.2; 3 is from 1 by possibility rules; 4 is from 2 by necessity rules; 5 is from 3 and 4 by the substitutivity rule; the branch now closes by 5.

2. This proof is very similar to the proof of 1. □

B.3.4 Decidability and Termination

Now the related issues of the decidability of the proof system and termination of the computation system are addressed.

First, the proof system is discussed. Many *propositional* sublogics of the logic considered here are decidable; for example, the fusion logics $\mathbf{K}_m$, $\mathbf{T}_m$, $\mathbf{K4}_m$, $\mathbf{S4}_m$, $\mathbf{S5}_m$, and $\mathbf{KD45}_m$ ($m \geq 1$) are all decidable. For such logics, interest naturally turns to the complexity of their decision procedure. For example, the satisfiability problems of $\mathbf{S5}_m$ and

1	$\neg\Box_{j_1}\cdots\Box_{j_r}(s=t)$	1.
1	$\Box_{j_1}\cdots\Box_{j_r}((s=t)=\top)$	2.
$1.1_{j_1}\ldots1_{j_r}$	$\neg(s=t)$	3.
$1.1_{j_1}\ldots1_{j_r}$	$((s=t)=\top)$	4.
$1.1_{j_1}\ldots1_{j_r}$	$\neg\top$	5.

Figure B.16: Proof of rank 1 of $\Box_{j_1}\cdots\Box_{j_r}(s=t)$

KD45$_m$ ($m \geq 2$) are PSPACE-complete; thus the corresponding decision problems are also PSPACE-complete.

However, even propositional modal logics can be undecidable; for example, product logics of dimension three or more are often undecidable. In any case, for the class of intended applications, the higher-order (or, at least, first-order) versions of the logic are needed, and these are undecidable. As Church proved, the validity problem of (classical) first-order logic is undecidable. Furthermore, Gödel proved that the validity problem of (classical) second-order logic is not semi-decidable. Gödel's result depends upon the use of standard models. It was Henkin who realized that by extending the class of models to so-called Henkin (also called general) models the collection of valid formulas could be reduced and a completeness result could be obtained for higher-order (and therefore second-order) logic. Furthermore, the compactness and Löwenheim-Skolem theorems can be proved in this setting. Note that, at this point, Lindström's (first) theorem can be applied to show that higher-order logic (with the Henkin semantics) is essentially just a variant of first-order logic. In spite of this result, something important has been gained: instead of being forced to express certain things awkwardly in first-order logic, the greater expressive power of higher-order logic can be used. This book contains plenty of illustrations of the advantages of the greater expressive power of higher-order logic.

So how does one cope with the undecidability of the decision problem of the logic? The answer is that for each application the theorem-proving tasks that will arise in that application have to be examined and shown to terminate (or else the application has to be redesigned and re-implemented in order to achieve this). What is pertinent here is that each application is a substantial system and that it makes sense to expend effort on ensuring the desired termination, in the same way as it makes sense to use conventional software engineering techniques to ensure the correctness of the software that implements an application. Furthermore, this suggestion is feasible: for a particular application the range of theorem-proving tasks (even for an adaptive system) is likely to be tightly circumscribed, so it should be possible to provide termination proofs for the theorem-proving tasks that arise.

Some design decisions help to ease the problem of undecidability. To begin with, (full and, therefore, undecidable) higher-order unification is avoided altogether for a decidable form of matching that is quite powerful enough for most applications. Also, experience with a variety of applications indicates that the crucial facility offered by the logic is computation, as in Section B.3.1. That is, the most common reasoning task is to evaluate a function given some arguments; in many cases, the use of the theorem prover is subsidiary to the computation and is sometimes even essentially propositional and decidable.

This leads to the issue of the termination of the computation component of the rea-

soning system. The computation component in effect provides a declarative programming language. Even though the logic is modal, programming in this language is no more complicated than programming in, say, Haskell. Programs have to be written, modal or not, and the usual concerns of writing programs that efficiently compute what is needed and, especially, avoid infinite loops are much the same as for less expressive languages. So the answer to the problem of non-termination is that, for a particular application, the programming tasks that will arise in that application have to be analysed and shown to terminate (or else the application has to be redesigned and re-implemented in order to achieve this). There are well-established techniques for doing this such as the use of well-founded orderings.

In summary, for logics as expressive as the one proposed in this book there is no hope for *general* decidability and termination results. Instead, for each application, it is necessary to show that the proof and computation tasks that arise in the application do indeed terminate; in the case where a proof or computation is intrinsically potentially non-terminating, resource-bounded techniques have to be brought to bear.

B.4 Structural Induction

This section contains useful material on well-founded sets.

B.4.1 Principle of Structural Induction

A *strict partial order* on a set A is a binary relation $<$ on A such that, for each $a, b, c \in A$, $a \not< a$ (irreflexivity), $a < b$ implies $b \not< a$ (asymmetry), and $a < b$ and $b < c$ implies $a < c$ (transitivity).

Suppose now that $<$ is a strict partial order on a set A. Then $<$ is a *well-founded order* if there is no infinite sequence $a_1, a_2, \ldots$ such that $a_{i+1} < a_i$, for $i \in \mathbb{N}$.

By way of an example, if Σ any alphabet, then the set Σ^* of all strings over Σ with the substring relation $\prec$ (that is, $s_1 \prec s_2$ iff s_1 is a proper substring of s_2) is a well-founded set.

Let A be a set with a strict partial order $<$ and $X \subseteq A$. An element $a \in X$ is *minimal* for X if $x \not< a$, for all $x \in X$.

Proposition B.4.1. *Let A be a set with a strict partial order. Then A is a well-founded set iff every nonempty subset X of A has a minimal element (in X).*

Proof. Straightforward. □

Proposition B.4.2. *Let A be a set with a well-founded order $<$. Let X be a subset of A satisfying the condition: for all $a \in A$, whenever $b \in X$, for all $b < a$, it follows that $a \in X$. Then $X = A$.*

Proof. Suppose that $X \neq A$. Thus $A \setminus X \neq \emptyset$ and so $A \setminus X$ has a minimal element a, say. Consider an element $b \in A$ such that $b < a$. By the minimality of a, it follows that $b \notin A \setminus X$ and thus $b \in X$. Since this is true for all $b < a$, it follows that $a \in X$, by the condition satisfied by X. This gives a contradiction and so $X = A$. □

The condition in Proposition B.4.2 satisfied by X implies that X must contain the minimal elements of A (since these have no predecessors).

Now here is the first principle of inductive construction on well-founded sets.

Proposition B.4.3. *Let A be a well-founded set and S a set. Then there exists a unique function $f : A \to S$ having arbitrary given values on the minimal elements of A and satisfying the condition that there is a rule that, for all $a \in A$, uniquely determines the value of $f(a)$ from the values $f(b)$, for $b < a$.*

Proof. Uniqueness is shown first. Suppose that there exist two distinct functions f and g having the same values on the minimal elements of A and satisfying the condition in the statement of the proposition. Let X be the set of elements of A on which f and g differ. Let a be a minimal element of X. Now a cannot be minimal in A because f and g agree on the minimal elements of A. Thus there exist elements $b \in A$ such that $b < a$. For such an element b, $f(b) = g(b)$, since $b \notin X$. By the condition satisfied by f and g, it follows that $f(a) = g(a)$, which gives a contradiction.

Next existence is demonstrated. Denote the set $\{b \in A \mid b \leq a\}$ by W_a. Let $X = \{a \in A \mid$ there exists a function f_a defined on W_a having the given values on the minimal elements of A in W_a and satisfying the uniqueness condition on $W_a\}$. I show by the induction principle of Proposition B.4.2 that $X = A$. Note that, if $a, b \in X$ and $b < a$, by the uniqueness part of the proof applied to W_b, it follows that $f_b(x) = f_a(x)$, for all $x \in W_b$. Suppose now that $a \in A$ and $b \in X$, for all $b < a$. By the preceding remark, $a \in X$ – it suffices to define f_a by $f_a(b) = f_b(b)$, for all $b < a$, and let $f_a(a)$ be the value uniquely determined by the rule. By Proposition B.4.2, $X = A$. Now define f by $f(a) = f_a(a)$, for all $a \in A$. Clearly, f has the required properties. □

Because of the type system used by the logic, there is a requirement for another principle of inductive construction that demonstrates the existence of functions that satisfy an extra condition. First a definition is needed.

Definition B.4.1. Let $\{A_i\}_{i \in I}$ be a partition of the set A, $\{S_i\}_{i \in I}$ a partition of the set S, and $f : A \to S$ a mapping. Then $a \in A$ is *consistent* (with respect to f, $\{A\}_{i \in I}$ and $\{S\}_{i \in I}$) if $a \in A_i$ implies $f(a) \in S_i$.

Here is the second principle of inductive construction on well-founded sets.

Proposition B.4.4. *Let A be a well-founded set, $\{A_i\}_{i \in I}$ a partition of A, S a set, and $\{S_i\}_{i \in I}$ a partition of S. Then there exists a unique function $f : A \to S$ such that $f(A_i) \subseteq S_i$, for all $i \in I$, satisfying the following conditions.*

1. *For all $a \in A$, if a is minimal, then a is consistent.*
2. *There is a rule that, for all $a \in A$, uniquely determines the value of $f(a)$ from the values $f(b)$, for $b < a$, in such a way that if each b is consistent, then a is consistent.*

Proof. The existence and uniqueness follows from Proposition B.4.3. It remains to show that $f(A_i) \subseteq S_i$, for all $i \in I$.

Let $X = \{a \in A \mid a$ is consistent$\}$. It suffices to show that $X = A$. Now X contains the minimal elements, by assumption. Suppose now that $a \in A$ and, for all $b < a$, $b \in X$. By assumption, $a \in X$. Thus, by Proposition B.4.2, $A = X$. □

Bibliographical Notes

In the past, the motivations for employing modal higher-order logic have mostly been of a philosophical or linguistic nature, and outside these areas there have been very few works. For a brief historical account of these motivations, the reader is referred to the handbook chapter of Muskens [116] which develops a logic that repairs some deficiencies in Montague's Intensional Logic. A recent account of modal higher-order logic, motivated by philosophical considerations (Gödel's ontological argument), is given in [50]. An earlier work motivated by mainly linguistic considerations is [54]. All these treatments differ quite markedly from the one in this book as none has a type system that is as rich as the one presented here, they are not concerned with computation (only, at most, deduction), and their intended applications are quite different (therefore, they emphasize different technical machinery).

This material of this book has benefitted greatly from the extensive literature on modal logic. In particular, the books [50], [51], and [53] were most helpful. Closely related are works on the application of modal logic to artificial intelligence of which [47] influenced parts of this book. The account of higher-order logic in [6] especially influenced the presentation here.

The advantages of using a higher-order approach to computational logic have been advocated for at least the last 30 years. First, the functional programming community has used higher-order functions from the very beginning. The latest versions of functional languages, such as Haskell [129], show the power and elegance of higher-order functions, as well as related features such as strong type systems. Of course, the traditional foundation for functional programming languages has been the λ-calculus, rather than a higher-order *logic*. However, it is possible to regard functional programs as equational theories in a logic such as the one introduced here and this also provides a useful semantics. The reasoning system presented in Appendix B.3 extends the core computational mechanisms of existing functional programming languages in that it also contains a theorem-proving system, it is modal, and it admits logic programming idioms through programming with abstractions.

In the 1980s, higher-order programming in the logic programming community was introduced through the language λProlog [117]. The logical foundations of λProlog are provided by almost exactly the logic studied here (with the modal facilities removed). However, a different sublogic is used for λProlog programs than the equational theories proposed here. In λProlog, program statements are higher-order hereditary Harrop formulas, a generalisation of the definite clauses used by Prolog. The language provides an elegant use of λ-terms as data structures, meta-programming facilities, universal quantification and implications in goals, amongst other features.

A long-term interest amongst researchers in declarative programming has been the goal of building integrated functional logic programming languages. Probably the best developed of these functional logic languages is the Curry language [72], which is the result of an international collaboration over the last two decades. A survey of functional logic programming up to 1994 is in [70]. A more recent survey, concentrating on Curry, can be found in [71].

There are many other outstanding examples of systems that exploit the power of higher-order logic. For example, the HOL system [60] is an environment for interactive theorem proving in higher-order logic. Its most outstanding feature is its high degree of

programmability through the meta-language ML. The system has a wide variety of uses from formalising pure mathematics to verification of industrial hardware. In addition, there are at least a dozen other systems related to HOL. On the theoretical side, much of the research in theoretical computer science, especially semantics, is based on the λ-calculus and hence is intrinsically higher order in nature.

The form of polymorphism employed in this appendix has a long history. In 1958, the concept of a polymorphic type appeared in combinatory logic [33], being called there the 'functional character' of an expression. In 1969, Hindley [76] introduced the idea of a principal type schema, which is the most general polymorphic type of an expression. Nearly ten years later, in a highly influential paper [108], Milner independently rediscovered these ideas and also introduced the first practical polymorphic type checker. The Hindley–Milner type system, as it has become known, has been widely used in functional and logic programming languages.

Of the specific constants in $\mathfrak{C}$ introduced on page 419, only $=_\alpha$ is truly fundamental as the other constants can be defined using it. The details of this can be found in [6]. The modal operators also can be defined away using a natural embedding of modal higher-order logic into higher-order logic (at least for the case when just modal formulas rather than the more general modal terms are admitted) [11]. This means that the modal part of the logic does not add any extra theoretical capability, but it does increase the expressive power of higher-order logic and, since this is important in practice, the modal form of the logic is used throughout.

A much fuller treatment of basic terms and further illustrations of their use are given in [96]. The presentation here of predicate rewrite systems extends the one in [96] to the modal case. Examples of the use of (non-modal) predicate rewrite systems for machine learning applications are given in [96]. Definition B.1.51 extends that of a (non-modal) standard predicate in [96], precisely in that the definition here allows modalities to appear. Considerably more detail about polymorphism (in the non-modal case) is given in [96].

This account of higher-order logic employs the constant domain semantics, in contrast to the varying domain semantics that is explained in [51]. In [51, p.105-107], it is shown that each semantics can simulate the other. The Barcan and converse Barcan formulas are discussed in [51, p.108]. For the varying domain semantics, the Barcan biterm fails to hold generally, even in the case of formulas. The idea for introducing local and global assumptions comes from [51].

Modal computation has been studied for 20 years or so, mostly in the logic programming community in the context of epistemic or temporal logic programming languages. Useful surveys of this work are in [127] and [56]. A recent paper showing the current state of the art of modal logic programming is [123]. What is common between these works and this book is the emphasis on epistemic and temporal modalities. What is different is that almost all are based on Prolog and are, therefore, first order, and it seems they either provide epistemic modalities or temporal modalities, but not both.

The reasoning system presented in this book extends typical tableau modal theorem proving systems, such as those in [50] and [51] that heavily influenced the presentation here, largely in that it also has a computational component. The idea of having a specialized equational reasoning component in a theorem-proving system is, of course, an old one. What is interesting here is that for agent applications it is the computational (equational reasoning) component that is the most important; the most common reasoning task in an

agent is to evaluate a function call. Generally, the theorem-proving tasks are subsidiary to this.

A non-modal version of the standard equality theory is given in [96]. Programming with abstractions is discussed in [96]. Other examples of (non-modal) computation can be found in [95] and [96]. The Haskell programming language is described in [129].

The decidability of the (propositional) fusion logics $\mathbf{K}_m$, $\mathbf{T}_m$, $\mathbf{K4}_m$, $\mathbf{S4}_m$, $\mathbf{S5}_m$, and $\mathbf{KD45}_m$ ($m \geq 1$) is proved in [47] That the satisfiability problems of $\mathbf{S5}_m$ and $\mathbf{KD45}_m$ ($m \geq 2$) are PSPACE-complete and thus the corresponding decision problems are also PSPACE-complete is discussed in [47]. The undecidability of some product logics of dimension three or more is proved in [53]. The compactness and Löwenheim-Skolem theorems are discussed in [50]. Lindström's (first) theorem is proved in [45, Ch. XII]. Proof of termination using well-founded orderings was studied in [38]

Well-founded induction is also known as Noetherian induction, which is named after Emmy Noether who made important contributions in algebra and physics [168].

Exercises

B.1 Prove that each subterm is a term.

B.2 Prove that a variable is free in a term iff it has a free occurrence in the term.

B.3 Prove Proposition B.1.7.

B.4 Prove Proposition B.1.23.

B.5 Prove Parts 8, 12, and 13 of Proposition B.2.1.

B.6 Prove Part 2 of Proposition B.2.8.

B.7 Prove Part 3 of Proposition B.2.32.

B.8 Complete the missing parts of the proof of Proposition B.3.4.

B.9 Prove Part 2 of Proposition B.3.12.

B.10 Prove Proposition B.4.1.

References

[1] Autonomous weapons: An open letter from AI and robotics researchers. Future of Life Institute, 2015. `http://futureoflife.org/open-letter-autonomous-weapons/`.

[2] Machine learning: The power and promise of computers that learn by example. The Royal Society, 2017. `https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf`.

[3] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[4] C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis*. Springer, third edition, 2006.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety, 2016. arXiv:1606.06565v2.

[6] P.B. Andrews. *An Introduction to Mathematical Logic and Type Theory: To Truth Through Proof*. Kluwer Academic Publishers, second edition, 2002.

[7] A. Bain and D. Crisan. *Fundamentals of Stochastic Filtering*. Springer, 2009.

[8] M. Baldoni. *Normal Multimodal Logics: Automatic Deduction and Logic Programming Extension*. PhD thesis, Università degli Studi di Torino, 1998.

[9] J. Barrat. *Our Final Invention*. Thomas Dunne Books, 2013.

[10] T. Bengtsson, P. Bickel, and B. Li. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. *IMS Collections. Probability and Statistics: Essays in Honor of David A. Freedman*, 2:316–334, 2008.

[11] C. Benzmüller and B. Woltzenlogel Paleo. Higher-order modal logics: Automation and applications. In W. Faber and A. Paschke, editors, *Reasoning Web 2015, LNCS 9203*, pages 32–74. Springer, 2015.

[12] P. Bickel, B. Li, and T. Bengtsson. Sharp failure rates for the bootstrap particle filter in high dimensions. *IMS Collections. Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, 3:318–329, 2008.

[13] P. Billingsley. *Probability and Measure*. Wiley, third edition, 1995.

[14] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[15] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

[16] M.A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, second edition, 2004.

[17] M.A. Boden. Computer models of creativity. *AI Magazine*, pages 23–34, Fall 2009.

[18] G. Boole. *An Investigation of the Laws of Thought on which are founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberly, 1854.

[19] G. Boole. *Studies in Logic and Probability*. Watts & Co, 1952.

[20] E.G. Boring. Intelligence as the tests test it. *New Republic*, 35:35–37, 1923.

[21] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[22] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI'98*, pages 33–42, 1998.

[23] X. Boyen and D. Koller. Exploiting the architecture of dynamic systems. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pages 313–320, 1999.

[24] E. Çinlar. *Probability and Stochastics*. Springer, 2011.

[25] D. Chen, S. Yang-Zhao, J.W. Lloyd, and K.S. Ng. Factored conditional filtering: Tracking states and estimating parameters in high-dimensional spaces. In preparation, 2022.

[26] A. Church. A formulation of the simple theory of types. *Journal of Symbolic Logic*, 5:56–68, 1940.

[27] S. Colton and G. Wiggins. Computational creativity: The final frontier? In *ECAI'12 Proceedings of the 20th European Conference on Artificial Intelligence*, pages 21–26. IOS Press, 2012.

[28] D. Cope. *Computer Models of Musical Creativity*. MIT Press, 2005.

[29] K. Crawford. *Atlas of AI*. Yale University Press, 2021.

[30] D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practioners. *IEEE Transactions on Signal Processing*, 50(3):736–746, 2002.

[31] D. Crisan and J. Míguez. Uniform convergence over time of a nested particle filtering scheme for recursive parameter estimation in state-space Markov models. *Advances in Applied Probability*, 49(4):1170–1200, 2017.

[32] D. Crisan and J. Míguez. Nested particle filters for online parameter estimation in discrete-time state-space Markov models. *Bernoulli*, 24(4A):3039–3086, 2018.

[33] H.B. Curry and R. Feys. *Combinatory Logic*. North-Holland, 1958.

[34] S. Das, D. Lawless, B. Ng, and A. Pfeffer. Factored particle filtering for data fusion and situation assessment in urban environments. In *Proceedings of the Seventh International Conference on Information Fusion*, pages 955–962, 2005.

[35] A. Davies, P. Veličković, L. Buesing, et al. Advancing mathematics by guiding human intuition with AI. *Nature*, 600:70–74, 2021. https://doi.org/10.1038/s41586-021-04086-x.

[36] L. De Raedt and K. Kersting. Probabilistic logic learning. *SIGKDD Explorations*, 5(1):31–48, 2003.

[37] L. Demey, B. Kooi, and J. Sack. Logic and probability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition, 2017.

[38] N. Dershowitz and Z. Manna. Proving termination with multiset orderings. *Communications of the ACM*, 22(8):465–476, 1979.

[39] Free dictionary, 2017. [Online; accessed 24-August-2017].

[40] P. Djurić and M. Bugallo. Particle filtering for high-dimensional systems. In *5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, CAMSAP 2013*, pages 352–355, 2013.

[41] P. Djurić, T. Lu, and M. Bugallo. Multiple particle filtering. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007*, pages 1181–1184, 2007.

[42] A. Doucet and A.M. Johansen. Tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovskii, editors, *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press, 2011.

[43] R.M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

[44] J. Dugundji. *Topology*. Allyn and Bacon, 1966.

[45] H.D. Ebbinghaus, J. Flum, and W. Thomas. *Mathematical Logic*. Springer-Verlag, 1984.

[46] R. Fagin and J.Y. Halpern. Reasoning about knowledge and probability. *Journal of the ACM*, 41(2):340–367, 1994.

[47] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

[48] W.M. Farmer. The seven virtues of simple type theory. *Journal of Applied Logic*, 6(3):267–286, 2008.

[49] D. Ferrucci et al. Building Watson: An overview of the DeepQA Project. *AI Magazine*, 31(3):59–79, 2010.

[50] M. Fitting. *Types, Tableaus, and Gödel's God*. Kluwer Academic Publishers, 2002.

[51] M. Fitting and R.L. Mendelsohn. *First-order Modal Logic*. Kluwer Academic Publishers, 1998.

[52] C. Frey and M. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017.

[53] D.M. Gabbay, A. Kurucz, F. Wolter, and M. Zakharyaschev. *Many-Dimensional Modal Logics: Theory and Applications*. Studies in Logic and The Foundations of Mathematics, Volume 148. Elsevier, 2003.

[54] D. Gallin. *Intensional and Higher-order Modal Logic*. North-Holland, 1975.

[55] S. Ganzfried. Reflections on the first man versus machine no-limit Texas Hold'em competition. *AI Magazine*, 38(2):77–85, 2017.

[56] M. Gergatsoulis. Temporal and modal logic programming languages. In A. Kent and J.G. Williams, editors, *Encyclopedia of Microcomputers*, volume 27, supplement 6, pages 393–408. Marcel Dekker, 2001.

[57] L. Getoor and B. Taskar, editors. *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[58] I.J. Good. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6:31–88, 1966.

[59] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[60] M.J.C. Gordon and T.F. Melham. *Introduction to HOL: A Theorem Proving Environment for Higher Order Logic*. Cambridge University Press, 1993.

[61] N.J. Gordon, D.J. Salmond, and A.F.M Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Proc. Inst. Elec. Eng.*, 140(2):107–113, 1993.

[62] W.T. Gowers. Rough structure and classification. In N. Alon, J. Bourgain, A. Connes, M. Gromov, and V. Milman, editors, *Visions in Mathematics*, pages 79–117. Birkhäuser Verlag, 2000.

[63] R.L. Gregory, editor. *The Oxford Companion to the Mind*. Oxford University Press, 1998.

[64] T. Hailperin. *Sentential Probability Logic*. Lehigh University Press, 1996.

[65] A. Hájek. Probability, logic and probability logic. In L. Goble, editor, *The Blackwell Guide to Philosophical Logic*, chapter 16, pages 362–384. Blackwell, 2001.

[66] J.Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990.

[67] J.Y. Halpern. *Reasoning about Uncertainty*. MIT Press, 2003.

[68] D. Hambrick and A. Burgoyne. The difference between rationality and intelligence. *The New York Times*, September 2016. `https://nyti.ms/2cM5MDU/`.

[69] S. O. Hansson. Logic of belief revision. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition, 2017.

[70] M. Hanus. The integration of functions into logic programming: From theory to practice. *Journal of Logic Programming*, 19&20:583–628, 1994.

[71] M. Hanus. Functional logic programming: From theory to Curry. In *Programming Logics - Essays in Memory of Harald Ganzinger*, pages 123–168. Springer LNCS 7797, 2013.

[72] M. Hanus (ed.). Curry: An integrated functional logic language. `http://www.informatik.uni-kiel.de/~curry`.

[73] Y.N. Harari. *Homo Deus: A Brief History of Tomorrow*. Vintage, 2016.

[74] H.W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[75] P.M. Hill and J.W. Lloyd. *The Gödel Programming Language*. MIT Press, 1994. Logic Programming Series.

[76] R. Hindley. The principal type scheme of an object in combinatory logic. *Transactions of the American Mathematical Society*, 146:29–60, 1969.

[77] F.-H. Hsu. *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*. Princeton University Press, 2002.

[78] M. Hutter. *Universal Artificial Intelligence*. Texts in Theoretical Computer Science. Springer, 2005.

[79] M. Hutter, J.W. Lloyd, K.S. Ng, and W.T.B. Uther. Probabilities on sentences in an expressive logic. *Journal of Applied Logic*, 11:386–420, 2013.

[80] J. Ichikawa and M. Steup. The analysis of knowledge. In E. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016. `https://plato.stanford.edu/archives/win2016/entries/knowledge-analysis/`.

[81] J. Jacod and P. Protter. *Probability Essentials*. Springer, second edition, 2004.

[82] D. Kahneman. *Thinking, Fast and Slow*. Farrar Straus Giroux, 2011.

[83] O. Kallenberg. *Foundations of Modern Probability*. Springer, second edition, 2002.

[84] R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–46, 1960.

[85] A. Kechris. *Classical Descriptive Set Theory*. Springer, 1995.

[86] K. Kersting and L. De Raedt. Bayesian logic programming: Theory and tool. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[87] A. Klenke. *Probability Theory: A Comprehensive Course*. Springer, second edition, 2014.

[88] D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT Press, 2009.

[89] A. Kolmogorov. *Foundations of the Theory of Probability*. Springer, 1933.

[90] B. Lake, T. Ullman, J. Tenenbaum, and S. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. `https://doi.org/10.1017/S0140525X16001837`.

[91] S. Legg and M. Hutter. A collection of definitions of intelligence. In B. Goertzel and P. Wang, editors, *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, volume 157 of *Frontiers in Artificial Intelligence and Applications*, pages 17–24, Amsterdam, NL, 2007. IOS Press.

[92] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

[93] D. Leivant. Higher-order logic. In D.M. Gabbay, C.J. Hogger, J.A. Robinson, and J. Siekmann, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 2, pages 230–321. Oxford University Press, 1994.

[94] J.W. Lloyd. *Foundations of Logic Programming*. Springer, second edition, 1987.

[95] J.W. Lloyd. Programming in an integrated functional and logic language. *Journal of Functional and Logic Programming*, 1999(3), March 1999.

[96] J.W. Lloyd. *Logic for Learning: Learning Comprehensible Theories from Structured Data*. Cognitive Technologies. Springer, 2003.

[97] J.W. Lloyd. Knowledge representation and reasoning in modal higher-order logic. `http://users.rsise.anu.edu.au/~jwl`, 2007.

[98] J.W. Lloyd and K.S. Ng. Learning modal theories. In S. Muggleton, R. Otero, and A. Tamaddoni-Nezhad, editors, *Proceedings of the 16th International Conference on Inductive Logic Programming (ILP 2006)*, pages 320–334. Springer, LNAI 4455, 2007.

[99] J.W. Lloyd and K.S. Ng. Probabilistic and logical beliefs. In M. Dastani, J. Leite, A. El Fallah Seghrouchni, and P. Torroni, editors, *Languages, Methodologies and Development Tools for Multi-Agent Systems, International Workshop, LADS 2007*, LNAI 5118, pages 19–36. Springer, 2008.

[100] J.W. Lloyd and K.S. Ng. Reflections on agent beliefs. In M. Baldoni, T. C. Son, M. B. van Riemsdijk, and M. Winikoff, editors, *Declarative Agent Languages and Technologies V, Fifth International Workshop, DALT 2007*, LNAI 4897, pages 122–139. Springer, 2008.

[101] J.W. Lloyd and K.S. Ng. Declarative programming for agent applications. *Autonomous Agents and Multi-Agent Systems*, 23:224–272, 2011. DOI: 10.1007/s10458-010-9138-1.

[102] R. Mallah. The landscape of AI safety and beneficence research: Input for brainstorming at Beneficial AI 2017. `https://futureoflife.org/landscape/ResearchLandscapeExtended.pdf?x57718`, 2017.

[103] G. Marcus and E. Davis. *Rebooting AI: Building Artificial Intelliegnce We Can Trust*. Pantheon Books, 2019.

[104] J. McCormack and Mark d'Inverno, editors. *Computers and Creativity*. Springer, 2012.

[105] J.-J.Ch. Meyer and W. van der Hoek. *Epistemic Logic for Computer Science and Artificial Intelligence*, volume 41 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1995.

[106] B. Milch, B. Marthi, S. Russell, D. Sontag, D.L. Ong, and A. Kolobov. BLOG: Probabilistic models with unknown objects. In L.P. Kaelbling and A. Saffiotti, editors, *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1352–1359, 2005.

[107] B. Milch and S. Russell. First-order probabilistic languages: Into the unknown. In S. Muggleton, R. Otero, and A. Tamaddoni-Nezhad, editors, *Inductive Logic Programming: 16th International Conference, ILP 2006*, pages 10–24. Springer, LNAI 4455, 2007.

[108] R. Milner. A theory of type polymorphism in programming. *Journal of Computer and System Sciences*, 17:348–375, 1978.

[109] T.M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[110] V. Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[111] M. Montemerlo, S. Thrun, and W. Whittaker. Conditional particle filters for simultaneous mobile robot localization and people-tracking. In *IEEE International Conference on Robotics and Automation*, pages 695–701, 2002.

[112] S. Muggleton. Stochastic logic programs. In L. De Raedt, editor, *Advances in Inductive Logic Programming*, pages 254–264. IOS Press, 1996.

[113] K. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[114] K. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2022.

[115] K. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[116] R. Muskens. Higher order modal logic. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, Studies in Logic and Practical Reasoning, pages 621–653. Elsevier, 2006.

[117] G. Nadathur and D.A. Miller. Higher-order logic programming. In D.M. Gabbay, C.J. Hogger, and J.A. Robinson, editors, *The Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 5, pages 499–590. Oxford University Press, 1998.

[118] M. Newman. *Networks*. Oxford University Press, second edition, 2018.

[119] B. Ng, L. Peshkin, and A. Pfeffer. Factored particles for scalable monitoring. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI'02*, pages 370–377, 2002.

[120] K.S. Ng. *Learning Comprehensible Theories from Structured Data*. PhD thesis, Computer Sciences Laboratory, The Australian National University, 2005.

[121] K.S. Ng and J.W. Lloyd. Probabilistic reasoning in a classical logic. *Journal of Applied Logic*, 7(2):218–238, 2009. DOI:10.1016/j.jal.2007.11.008.

[122] K.S. Ng, J.W. Lloyd, and W.T.B. Uther. Probabilistic modelling, inference and learning using logical theories. *Annals of Mathematics and Artificial Intelligence*, 54:159–205, 2008. DOI:10.1007/s10472-009-9136-7.

[123] L.A. Nguyen. Multimodal logic programming. *Theoretical Computer Science*, 360:247–288, 2006.

[124] N.J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28(1):71–88, 1986.

[125] N.J. Nilsson. *Understanding Beliefs*. MIT Press, 2014.

[126] C. Nowzari, V.M. Preciado, and G.J. Pappas. Analysis and control of epidemics: A survey of spreading processes on complex networks. *IEEE Control Systems Magazine*, 36(1):26–46, 2016.

[127] M.A. Orgun and W. Ma. An overview of temporal and modal logic programming. In D.M. Gabbay and H.J. Ohlbach, editors, *Proceedings of the First International Conference on Temporal Logics (ICTL'94)*, volume 827 of *Lecture Notes in Artificial Intelligence*, pages 445–479. Springer, 1994.

[128] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of Modern Physics*, 87(3), 2015.

[129] S. Peyton Jones. *Haskell 98 Language and Libraries*. Cambridge University Press, 2003.

[130] S. Peyton Jones and J. Hughes (editors). Haskell98: A non-strict purely functional language. `http://haskell.org/`.

[131] J. Piaget and M.T. Cook. *The Origins of Intelligence in Children*. International University Press, 1952.

[132] D. Poole. First-order probabilistic inference. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 985–991, 2003.

[133] P. Rebeschini. *Nonlinear filtering in high dimension*. PhD thesis, Princeton University, 2014.

[134] P. Rebeschini and R. Van Handel. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability*, 25(5):2809–2866, 2015.

[135] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

[136] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, second edition, 2004.

[137] S. Russell. Unifying logic and probability. *Communications of the ACM*, 58(7):88–97, 2015.

[138] S. Russell. *Human Compatible: AI and the Problem of Control*. Penguin Books, 2019.

[139] S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI magazine*, 36:105–114, 2015.

[140] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, third edition, 2010.

[141] S. Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[142] S. Shapiro. Classical logic II – Higher-order logic. In L. Goble, editor, *The Blackwell Guide to Philosophical Logic*, pages 33–54. Blackwell, 2001.

[143] A. Shirazi and E. Amir. Probabilistic modal logic. In R.C. Holte and A. Howe, editors, *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 489–495, 2007.

[144] S. Shreve. *Stochastic Calculus for Finance II: Continuous-time Models*. Springer, 2004.

[145] D. Silver et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[146] D. Silver et al. Mastering the game of Go without human knowledge. *Nature*, 550:354–359, 2017.

[147] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640, 2008.

[148] C. Snyder, T. Bengtsson, P. Bickel, and J.Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640, 2008.

[149] K.E. Stanovich. *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. The University of Chicago Press, 2004.

[150] K.E. Stanovich. *Rationality and the Reflective Mind*. Oxford University Press, 2011.

[151] K.E. Stanovich. On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In K. Holyoak and R. Morrison, editors, *The Oxford Handbook of Thinking and Reasoning*, pages 433–455. Oxford University Press, 2012.

[152] K.E. Stanovich, R.F. West, and M.E. Toplak. *The Rationality Quotient: Toward a Test of Rational Thinking*. The MIT Press, 2016.

[153] A. Svensson, T. Schön, and M. Kok. Nonlinear state space smoothing using the conditional particle filter. *IFAC-PapersOnLine*, 48:975–980, 2015.

[154] M. Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Allen Lane, 2017.

[155] P. Tetlock and B. Mellers. The great rationality debate. *Psychological Science*, 13:94–99, 2002.

[156] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[157] A. Tversky and D. Kahneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

[158] J. van Benthem and K. Doets. Higher-order logic. In D.M. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic*, volume 1, pages 275–330. Reidel, 1983.

[159] P. van Leeuwen. Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12):4089–4114, 2009.

[160] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[161] J. von Neumann and O. Morgenstern. *The Theory of Games and Economic Behaviour*. Princeton University Press, 1944.

[162] T. Walsh. *2062: The World that AI Made*. La Trobe University Press, 2018.

[163] L. Wasserman. *All of Statistics*. Texts in Statistics. Springer, 2004.

[164] Wikipedia. Artificial general intelligence – Wikipedia, the free encyclopedia, 2017. [Online; accessed 6-August-2017].

[165] Wikipedia. Automated theorem proving – Wikipedia, the free encyclopedia, 2017. [Online; accessed 6-August-2017].

[166] Wikipedia. Empiricism – Wikipedia, the free encyclopedia, 2017. [Online; accessed 17-August-2017].

[167] Wikipedia. Belief revision – Wikipedia, the free encyclopedia, 2018. [Online; accessed 8-February-2018].

[168] Wikipedia. Emmy Noether – Wikipedia, the free encyclopedia, 2018. [Online; accessed 4-April-2018].

[169] Wikipedia. Fifth generation computer – Wikipedia, the free encyclopedia, 2018. [Online; accessed 26-April-2018].

[170] Wikipedia. Epistemology — Wikipedia, the free encyclopedia, 2019. [Online; accessed 3-April-2019].

[171] J. Williamson. Probability logic. In D. Gabbay, R. Johnson, H.J. Ohlbach, and J. Woods, editors, *Handbook of the Logic of Inference and Argument: The Turn Toward the Practical*, volume 1 of *Studies in Logic and Practical Reasoning*, pages 397–424. Elsevier, 2002.

[172] M. Wooldridge. *Reasoning about Rational Agents*. MIT Press, 2000.

[173] M. Wooldridge. *The Road to Conscious Machines: The Story of AI*. Pelican Books, 2020.

Index