Structured Probabilistic Reasoning

(Incomplete draft)

Bart Jacobs
Institute for Computing and Information Sciences, Radboud University Nijmegen,
P.O. Box 9010, 6500 GL Nijmegen, The Netherlands.
bart@cs.ru.nl http://www.cs.ru.nl/~bart

Version of: November 30, 2021

Contents

	Preface		*page* v
1	**Collections**		1
	1.1	Cartesian products	2
	1.2	Lists	6
	1.3	Subsets	16
	1.4	Multisets	25
	1.5	Multisets in summations	35
	1.6	Binomial coefficients of multisets	41
	1.7	Multichoose coefficents	48
	1.8	Channels	56
	1.9	The role of category theory	61
2	**Discrete probability distributions**		69
	2.1	Probability distributions	70
	2.2	Probabilistic channels	81
	2.3	Frequentist learning: from multisets to distributions	92
	2.4	Parallel products	98
	2.5	Projecting and copying	109
	2.6	A Bayesian network example	116
	2.7	Divergence between distributions	122
3	**Drawing from an urn**		127
	3.1	Accumlation and arrangement, revisited	132
	3.2	Zipping multisets	136
	3.3	The multinomial channel	144
	3.4	The hypergeometric and Pólya channels	156
	3.5	Iterated drawing from an urn	168
	3.6	The parallel multinomial law: four definitions	181
	3.7	The parallel multinomial law: basic properties	189

	3.8	Parallel multinomials as law of monads	200
	3.9	Ewens distributions	207
4	**Observables and validity**		221
	4.1	Validity	222
	4.2	The structure of observables	234
	4.3	Transformation of observables	249
	4.4	Validity and drawing	256
	4.5	Validity-based distances	264
5	**Variance and covariance**		275
	5.1	Variance and shared-state covariance	276
	5.2	Draw distributions and their (co)variances	284
	5.3	Joint-state covariance and correlation	293
	5.4	Independence for random variables	299
	5.5	The law of large numbers, in weak form	304
6	**Updating distributions**		311
	6.1	Update basics	312
	6.2	Updating draw distributions	322
	6.3	Forward and backward inference	328
	6.4	Discretisation, and coin bias learning	346
	6.5	Inference in Bayesian networks	354
	6.6	Bayesian inversion: the dagger of a channel	358
	6.7	Pearl's and Jeffrey's update rules	366
	6.8	Frequentist and Bayesian discrete probability	383
7	**Directed Graphical Models**		390
	7.1	String diagrams	392
	7.2	Equations for string diagrams	399
	7.3	Accessibility and joint states	404
	7.4	Hidden Markov models	408
	7.5	Disintegration	419
	7.6	Disintegration for states	428
	7.7	Disintegration and inversion in machine learning	438
	7.8	Categorical aspects of Bayesian inversion	450
	7.9	Factorisation of joint states	457
	7.10	Inference in Bayesian networks, reconsidered	466
	References		473

Preface

No victor believes in chance.

Friedrich Nietzsche,
Die fröhliche Wissenschaft, §258, 1882.

Originally in German:
Kein Sieger glaubt an den Zufall.

Probability is for losers — a defiant rephrase of the above aphorism of the German philosopher Friedrich Nietzsche. According to him, winners do not reason with probabilities, but with certainties, via Boolean logic, one could say. However, this goes against the current trend, in which reasoning with probabilities has become the norm, in large scale data analytics and in aritificial intelligence (AI); it is Boolean reasoning that is now losing influence, see also the famous *End of Theory* article [4] from 2008. This book is about the mathematical structures underlying the reasoning, not of Nietzsche's winners, but of today's apparent winners.

The phrase 'structure in probability' in the title of this book may sound like a *contradictio in terminis*: it seems that probability is about randomness, like in the tossing of coins, in which one may not expect to find much structure. Still, as we know since the seventeenth century, via the pioneering work of Christiaan Huygens, Pierre Fermat, and Blaise Pascal, there is quite some mathematical structure in the area of probability. The *raison d'être* of this book is that there is more structure — especially algebraic and categorical — than is commonly emphasised.

The scientific roots of this book's author lie outside probability theory, in type theory and logic (including some quantum logic), in semantics and specification of programming languages, in computer security and privacy, in state-based computation (coalgebra), and in category theory. This scientific distance to probability theory has advantages and disadvantages. Its obvious disadvan-

tage is that there is no deeply engrained familiarity with the field and with its development. But at the same time this distance may be an advantage, since it provides a fresh perspective, without sacred truths and without adherence to common practices and notations. For instance, the terminology and notation in this book are influenced by quantum theory, *e.g.* in using the words 'state', 'observable' and 'test' — as synonyms for 'distribution', for an $\mathbb{R}$-valued function on a sample space and for compatible (summable) predicates — in using ket notation $| - \rangle$ in writing discrete probability distributions, or in using daggers as reversals, in analogy with conjugate transposes (for Hilbert spaces).

It should be said: for someone trained in formal methods, the area of probability theory can be rather sloppy: everything is called 'P', types are hardly ever used, crucial ingredients (like distributions in expected values) are left implicit, basic notions (like conjugate prior) are introduced only via examples, calculation recipes and algorithms are regularly just given, without explanation, goal or justification, *etc.* This hurts, especially because there is so much beautiful mathematical structure around. For instance, the notion of a channel (see below) formalises the idea of a conditional probability and carries a rich mathematical structure that can be used in compositional reasoning, with both sequential and parallel composition. The Bayesian inversion ('dagger') of a channel does not only come with appealing mathematical (categorical) properties — *e.g.* smooth interaction with sequential and parallel composition — but is also extremely useful in inference and learning. Via this dagger we can connect forward and backward inference (see Theorem 6.6.3: backward inference is forward inference with the dagger, and vice-versa) and capture the difference between Pearl's and Jeffrey's update rules (see Theorem 6.7.4: Pearl increases validity, whereas Jeffrey decreases divergence).

We even dare to think that this 'sloppiness' is ultimately a hindrance to further development of the field, especially in computer science, where computer-assisted reasoning requires a clear syntax and semantics. For instance, it is hard to even express the above-mentioned theorems 6.6.3 and 6.7.4 in standard probabilistic notation. One can speculate that states/distributions are kept implicit in traditional probability theory because in many examples they are used as a fixed implicit assumption in the background. Indeed, in mathematical notation one tends to omit — for efficiency — the least relevant (implicit) parameters. But the essence of probabilistic computation is state transformation, where it has become highly relevant to know explicitly in which state one is working at which stage. The notation developed in this book helps in such situations — and in many other situations as well, we hope.

Apart from having beautiful structure, probability theory also has *magic*. It can be found in the following two points.

1 Probability distributions can be updated, making it possible to absorb information (evidence) into them and learn from it. Multiple updates, based on data, can be used for training, so that a distribution absorbs more and more information and can subsequently be used for prediction or classification.
2 The components of a joint distribution, over a product space, can 'listen' to each other, so that updating in one (product) component has crossover ripple effects in other components. This ripple effect looks like what happens in quantum physics, where measuring one part of an entangled quantum systems changes other parts.

The combination of these two points is very powerful and forms the basis for probabilistic reasoning. For instance, if we know that two phenomena are related, and we have new information about one of them, then we also learn something new about the other phenomenon, after updating. We shall see that such crossover ripple effects can be described in two equivalent ways, starting from a joint distribution with evidence in one component.

- We can use the "weaken-update-marginalise" approach, where we first weaken the evidence from one component to the whole product space, so that it fits the joint distribution and can be used for updating; subsequently, we marginalise the updated state to the component that we wish to learn more about. That's where the ripple effect through the joint distribution becomes visible.
- We can also use the "extract-infer" technique, where we first extract a conditional probability (channel) from the joint distribution and then do (forward or backward) inference with the evidence, along the channel. This is what happens if we reason in Bayesian networks, when information at one point in the network is transported up and down the connections to draw conclusions at another point in the network.

The equivalence of these two approaches will be demonstrated and exploited at various places in this book.

Here is a characteristic illustration of our structure-based approach. A well known property of the Poisson distribution *pois* is commonly expressed as: if $X_1 \sim pois[\lambda_1]$ and $X_2 \sim pois[\lambda_2]$ then $X_1 + X_2 \sim pois[\lambda_1 + \lambda_2]$. This formulation uses random variables X_1, X_2, which are Poisson-distributed. We shall formulate this fact as an (algebraic) structure preservation property of the Poisson distribution, without using any random variables. The property is: $pois[\lambda_1 + \lambda_2] = pois[\lambda_1] + pois[\lambda_2]$. It says that the Poisson channel *pois* is a is a homomorphism of monoids, from non-negative reals $\mathbb{R}_{\geq 0}$ to distributions $\mathcal{D}(\mathbb{N})$ on the natural numbers, see Proposition 2.4.6 for details. This result uses

a commutative monoid structure on distributions whose underlying space is itself a commutative monoid. This monoid structure plays a role in many other situations, for instance in the fundamental distributive law that turns multisets of distributions into distributions of multisets.

The following aspects characterise the approach of this book.

1 *Channels* are used as a cornerstone in probabilistic reasoning. The concept of (communication) channel is widely used elsewhere, under various names, such as conditional probability, stochastic matrix, probabilistic classifier, Markov kernel, conditional probability table (in Bayesian network), probabilistic function/computation, signal (in Bayesian persuasion theory), and finally as Kleisli map (in category theory). Channels can be composed sequentially and in parallel, and can transform both states and predicates. Channels exist for all relevant collection types (lists, subsets, multisets, distributions), for instance for non-deterministic, and for probabilistic computation. However, after the first chapter about collection types, channels will be used exclusively for distributions, in probabilistic form.
2 *Multisets* play a central role to capture various forms of data, like coloured balls in an urn, draws from such an urn, tables, inputs for successive learning steps, *etc.* The interplay between multisets and distributions, notably in learning, is a recurring theme.
3 *States* (distributions) are treated as separate from, and dual, to *predicates*. These predicates are the ingredients of an (implicit) probabilistic logic, with, for instance conjunction and negation operations. States are really different entities, with their own operations, without, for instance conjunction and negation. In this book, predicates standardly occur in fuzzy (soft, non-sharp) form, taking values in the unit interval $[0, 1]$. Along a channel one can transfer states forward, and predicates backward. A central notion is the validity of a predicate in a state, written as $\models$. It is standardly called expected value. Conditioning involves updating a state with a predicate.

It is not just mathematical aesthetics that drives the developments in this book. Probability theory nowadays forms the basis of large parts of big data analytics and of artificial intelligence. These areas are of increasing societal relevance and provide the basis of the modern view of the world — more based on correlation than on causation — and also provide the basis for much of modern decision making, that may affect the lives of billions of people in profound ways. There are increasing demands for justification of such probabilistic reasoning methods and decisions, for instance in the legal setting provided by Europe's General Data Protection Regulation (GDPR). Its recital 71 is about automated decision-making and talks about a right to obtain an explanation:

In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

It is not acceptable that your mortgage request is denied because you drive a blue car — in presence of a correlation between driving blue cars and being late on one's mortgage payments.

These and other developments have led to a new area called Explainable Artificial Intelligence (XAI), which strives to provide decisions with explanations that can be understood easily by humans, without bias or discrimination. Although this book will not contribute to XAI as such, it aims to provide a mathematically solid basis for such explanations.

In this context it is appropriate to quote Judea Pearl [134] from 1989 about a divide that is still wide today.

To those trained in traditional logics, symbolic reasoning is the standard, and nonmonotonicity a novelty. To students of probability, on the other hand, it is symbolic reasoning that is novel, not nonmonotonicity. Dealing with new facts that cause probabilities to change abruptly from very high values to very low values is a commonplace phenomenon in almost every probabilistic exercise and, naturally, has attracted special attention among probabilists. The new challenge for probabilists is to find ways of abstracting out the numerical character of high and low probabilities, and cast them in linguistic terms that reflect the natural process of accepting and retracting beliefs.

This book does not pretend to fill this gap. One of the big embarrassments of the field is that there is no widely accepted symbolic logic for probability, together with proof rules and a denotational semantics. Such a logic for symbolic reasoning about probability will be non-trivial, because it will have to be non-monotonic[1] — a property that many logicians shy away from. This book does aim to contribute towards bridging the divide mentioned by Pearl, by providing a mathematical basis for such a symbolic probabilistic logic, consisting of channels, states, predicates, transformations, conditioning, disintegration, *etc.*

From the perspective of this book, the structured categorical approach to probability theory began with the work of Bill Lawvere (already in the 1960s) and his student Michèle Giry. They recognised that taking probability distributions has the structure of a monad, which was published in the early 1980s in [58]. Roughly at the same time Dexter Kozen started the systematic investigation of probabilistic programming languages and logics, published in [106, 107]. The monad introduced back then is now called the Giry mo-

[1] Informally, a logic is non-monotonic if adding assumptions may make a conclusion less true. For instance, I may think that scientists are civilised people, until, at some conference dinner, a heated scientific debate ends in a fist fight.

nad $\mathcal{G}$, whose restriction to finite discrete probability distributions is written as $\mathcal{D}$. Much of this book, certainly in the beginning, concentrates on this discrete form. The language and notation that is used, however, covers both discrete and continuous probability — and quantum probability too (inspired by the general categorical notion of effectus, see [22, 68]).

Since the early 1980s the area of categorical probability theory remained relatively calm. It is only in the new millenium that there is renewed attention, sparked in particular by several developments.

- The grown interest in probabilistic programming languages that incorporate updating (conditioning) and/or higher order features, see *e.g.* [29, 30, 32, 129, 154, 153].
- The compositional approach to Bayesian networks [26, 48] and to Bayesian reasoning [28, 87, 89].
- The use of categorical and diagrammatic methods in quantum foundations, including quantum probability, see [25] for an overview.
- The efforts to develop 'synthetic' probability theory via a categorical axiomatisation, see *e.g.* [52, 53, 147, 20, 80].

This book builds on these developments.

The intended audience consists of students and professionals — in mathematics, computer science, artificial intelligence and related fields — with a basic background in probability and in algebra and logic — and with an interest in formal, logically oriented approaches. This book's goal is not to provide intuitive explanations of probability, like [158], but to provide clear and precise formalisations of the relevant structures. Mathematical abstraction (esp. categorical air guitar playing) is not a goal in itself (except maybe in Chapter 3): instead, the book tries to uncover relevant abstractions in concrete problems. It includes several basic algorithms, with a focus on the algorithms' correctness, not their efficiency. Each section ends with a series of exercises, so that the book can also be used for teaching and/or self-study. It aims at an undergraduate level. No familiarity with category theory is assumed. The basic, necessary notions are explained along the way. People who wish to learn more about category theory can use the references in the text, consult modern introductory texts like [7, 113], or use online resources such as ncatlab.org or Wikipedia.

Contents overview

The first chapter of the book covers introductory material that is meant to set the scene. It starts from basic collection types like lists and subsets, and continues with multisets, which receive most attention. The chapter discusses the

(free) monoid structure on all these collection types and introduces 'unit' and 'flatten' maps as their common, underlying structure. It also introduces the basic concept of a channel, for these three collection types, and shows how channels can be used for state transformation and how they can be composed, both sequentially and in parallel. At the end, the chapter provides definitions of the relevant notions from category theory.

In the second chapter (discrete) probability distributions first emerge, as a special collection type, with their own associated form of (probabilistic) channel. The subtleties of parallel products of distributions (states), with entwinedness/correlation between components and the non-naturality of copying, are discussed at this early stage. This culminates in an illustration of Bayesian networks in terms of (probabilistic) channels. It shows how predictions are made within such Bayesian networks via state transformation and via compositional reasoning, basically by translating the network structure into (sequential and parallel) composites of channels.

Blindly drawing coloured balls from an urn is a basic model in discrete probability. Such draws are analysed systematically in Chapter 3, not only for the two familar multinomial and hypergeometric forms (with or without replacement), but also in the less familiar Pólya and Ewens forms. By describing these draws as probabilistic channels we can derive the well known formulations for these draw-distributions via channel-composition. Once formulated in terms of channels, these distributions satisfy various compositionality properties. They are typical for our approach and are (largely) absent in traditional treatments of this topic. Urns and draws from urns are both described as multisets. The interplay between multisets and distributions is an underlying theme in this chapter. There is a fundamental distributive law between multisets and distributions that expresses basic structural properties.

The fourth chapter is more logically oriented, via observables $X \to \mathbb{R}$ (including factors, predicates and events) that can be defined on sample spaces X, providing numerical information. The chapter concentrates on validity of obervables in states and on transformation of observables. Where the second chapter introduces state transformation along a probabilistic channel in a forward direction, this fourth chapter adds observable (predicate) transformation in a backward direction. These two operations are of fundamental importance in program semantics, and also in quantum computation — where they are distinguished as Schrödinger's (forward) and Heisenberg's (backward) approach. In this context, a random variable is a combination of a state and an observable, on the same underlying sample space. The statistical notions of variance and covariance are described in terms of of validity for such random variables in

Chapter 5. This chapter distinguishes two forms of covariance, with a 'shared' or a 'joint' state, which satisfy different properties.

A very special technique in the area of probability theory is conditioning, also known as belief updating, or simply as updating. It involves the incorporation of evidence into a distribution (state), so that the distribution better fits the evidence. In traditional probability such conditioning is only indirectly available, via a rule $P(B \mid A)$ for computing conditional probabilities. In Chapter 6 we formulate conditioning as an explicit operation, mapping a state ω and a predicate p to a new updated state $\omega|_p$. A key result is that the validity of p in $\omega|_p$ is higher than the validity of p in the original state ω. This means that we have learned from p and adapted our state (of mind) from ω to $\omega|_p$. This updating operation $\omega|_p$ forms an action (of predicates on states) and satisfies Bayes' rule, in fuzzy form. The combination with forward and backward transformation along a channel leads to the techniques of forward inference (causal reasoning) and backward inference (evidential reasoning). These inference techniques are illustrated in many examples, including Bayesian networks. The backward inference rule is also called Pearl's rule. It turns out that there exists an alternative update mechanism, called Jeffrey's rule. It can give completely different outcomes. The formalisation of Jeffrey's rule is based on the reversal of the direction of a channel. This corresponds to turning a conditional probability $P(y \mid x)$ into $P(x \mid y)$, essentially via Bayes' rule. This Bayesian inversion is also called the 'dagger' of a channel, since it satisfies the properties of a dagger operation (or conjugate transpose), in Hilbert spaces (and in quantum theory). This chapter includes a mathematical characterisation of the different goals of these two update rules: Pearl's rule increases validity and Jeffrey's rule decreases divergence. More informally, one learns via Pearl's rule by improving what's going well and via Jeffrey's rule by reducing what's going wrong. Jeffrey's rule is thus an error correction mechanism. This fits the basic idea in predictive coding theory [64, 23] that the human mind is seen as a Bayesian prediction engine that operates by reducing prediction errors.

Since channels can be composed both sequentially and in parallel we can use graphical techniques to represent composites of channels. So-called string diagrams have been developed in physics and in category theory to deal with the relevant compositional structure (symmetric monoidal categories). Chapter 7 introduces these (directed) graphical techniques. It first describes string diagrams. They are similar to the graphs used for Bayesian networks, but they have explicit operations for copying and discarding and are thus more expressive. But most importantly, string diagrams have a clear semantics, namely in terms of channels. The chapter illustrates these string diagrams in a channel-based description of the basics of Markov chains and of hidden Markov mod-

els. But the most fundamental technique that is introduced in this chapter, via string diagrams, is disintegration. In essence, it is the well known procedure of extracting a conditional probability $P(y \mid x)$ from a joint probability $P(x, y)$. One of the themes running through this book is how 'crossover' influence can be captured via channels — extracted from joint states via disintegration — in particular via forward and backward inference. This phenomenon is what makes (reasoning in) Bayesian networks work. Disintegration is of interest in itself, but also provides an intuitive formalisation of the Bayesian inversion of a channel.

Almost all of the material in these chapters is known from the literature, but typically not in the channel-based form in which it is presented here. This book includes many examples, often copied from familiar sources, with the deliberate aim of illustrating how the channel-based approach actually works. Since many of these examples are taken from the literature, the interested reader may wish to compare the channel-based description used here with the original description.

Status of the current incomplete version

An incomplete version of this book is made available online, in order to generate feedback and to justify a pause in the writing process. Feedback is most welcome, both positive and negative, especially when it suggests concrete improvements of the text. This may lead to occasional updates of this text. The date on the title page indicates the current version.

Some additional points.

- The (non-trivial) calculations in this book have been carried out with the EfProb library [20] for channel-based probability. Several calculations in this book can be done by hand, typically when the outcomes are described as fractions, like $\frac{117}{2012}$. Such calculations are meant to be reconstructable by a motivated reader who really wishes to learn the 'mechanics' of the field. Doing such calculations is a great way to really understand the topic — and the approach of this book[2]. Outcomes written in decimal notation 0.1234, as approximations, or as plots, serve to give an impression of the results of a computation.
- For the rest of this book, beyond Chapter 7, several additional chapters exist

[2] Doing the actual calculations can be a bit boring and time consuming, but there are useful online tools for calculating fractions, such as https://www.mathpapa.com/fraction-calculator.html. Recent versions of EfProb also allow calculations in fractional form.

in unfinished form, for instance on learning, probabilistic automata, causality and on continuous probability. They will be incorporated in due course.

Bart Jacobs, August 26, 2021.

1

Collections

There are several ways to put elements from a given set together, for instance as lists, subsets, multisets, and as probability distributions. This introductory chapter takes a systematic look at such collections and seeks to bring out numerous similarities. For instance, lists, subsets and multisets all form monoids, by suitable unions of collections. Unions of distributions are more subtle and take the form of convex combinations. Also, subsets, multisets and distributions can be combined naturally via parallel products $\otimes$, though lists cannot. In this first chapter, we collect some basic operations and properties of tuples, lists, subsets and multisets — where multisets are 'sets' in which elements may occur multiple times. Probability distributions will be postponed to the next chapter. Especially, we collect several basic definitions and results for multisets, since they play an important role in the sequel, as urns, filled with coloured balls, as draws from such urns, and as data in learning.

The main differences between lists, subsets and multisets are summarised in the table below.

	lists	**subsets**	**multisets**
order of elements matters	+	-	-
multiplicity of elements matters	+	-	+

For instance, the lists $[a, a, b]$, $[a, b, a]$ and $[a, b]$ are all different. The multisets $2|a\rangle + 1|b\rangle$ and $1|b\rangle + 2|a\rangle$, with the element a occurring twice and the element b occurring once, are the same. However, $1|a\rangle + 1|b\rangle$ is a different multiset. Similarly, the sets $\{a, b\}$, $\{b, a\}$, and $\{a\} \cup \{a, b\}$ are the same.

These collections are important in themselves, in many ways, and primarily (in this book) as outputs of channels. Channels are functions of the form $\mathit{input} \to T(\mathit{output})$, where T is a 'collection' operator, for instance, combining elements as lists, subsets, multisets, or distributions. Such channels capture

a form of computation, directly linked to the form of collection that is produced on the outputs. For instance, channels where T is powerset are used as interpretations of non-deterministic computations, where each input element produces a subset of possible output elements. In the probabilistic case these channels produce distributions, with a suitable instantiation of the operator T. Channels can be used as elementary units of computation, which can be used to build more complicated computations via sequential and parallel composition.

First, the reader is exposed to some general considerations that set the scene for what is coming. This requires some level of patience. But it is useful to see the similarities between probability distributions (in Chapter 2) and other collections first, so that constructions, techniques, notation, terminology and intuition that we use for distributions can be put in a wider perspective and thus may become natural.

The final section of this chapter explains where the abstractions that we use come from, namely from category theory, especially referring the notion of monad. It gives a quick overview of the most relevant parts of this theory and also how category theory will be used in the remainder of this book to elicit mathematical structure. We use category theory pragmatically, as a tool, and not as a goal in itself.

1.1 Cartesian products

This section briefly reviews some (standard) terminology and notation related to Cartesian products of sets.

Let X_1 and X_2 be two arbitrary sets. We can form their Cartesian product $X_1 \times X_2$, as the new set containing all pairs of elements from X_1 and X_2, as in:

$$X_1 \times X_2 \coloneqq \{(x_1, x_2) \mid x_1 \in X_1 \text{ and } x_2 \in X_2\}.$$

(We use the sign $\coloneqq$ for mathematical definitions.)

We thus write (x_1, x_2) for the 'pair' or 'tuple' of elements $x_1 \in X_1$ and $x_2 \in X_2$. We have just defined a *binary* product set, constructed from two given sets X_1, X_2. We can also do this in n-ary form, for n sets $X_1, \ldots, X_n$. We then get an n-ary Cartesian product:

$$X_1 \times \cdots \times X_n \coloneqq \{(x_1, \ldots, x_n) \mid x_1 \in X_1, \ldots, x_n \in X_n\}.$$

The tuple $(x_1, \ldots, x_n)$ is sometimes called an n-tuple. For convenience, it may be abbreviated as a vector $\vec{x}$. The product $X_1 \times \cdots \times X_n$ is sometimes written

differently using the symbol $\prod$, as:

$$\prod_{1 \leq i \leq n} X_i \qquad \text{or more informally as:} \qquad \prod X_i.$$

In the latter case it is left implicit what the range is of the index element i.

We allow $n = 0$. The resulting 'empty' product is then written as a singleton set, written as 1, containing the empty tuple () as sole element, as in:

$$1 \coloneqq \{()\}.$$

For $n = 1$ the product $X_1 \times \cdots \times X_n$ is (isomorphic to) the set X_1. Note that we are overloading the symbol 1 and using it both as numeral and as singleton set.

If one of the sets X_i in a product $X_1 \times \cdots \times X_n$ is empty, then the whole product is empty. Also, if all of the sets X_i are finite, then so is the product $X_1 \times \cdots \times X_n$. In fact, the number of elements of $X_1 \times \cdots \times X_n$ is then obtained by multiplying all the numbers of elements of the sets X_i.

1.1.1 Projections and tuples

If we have sets $X_1, \ldots, X_n$ as above, then for each number i with $1 \leq i \leq n$ there is a projection function π_i out of the product to the set X_i, as in:

$$X_1 \times \cdots \times X_n \xrightarrow{\pi_i} X_i \qquad \text{given by} \qquad \pi_i(x_1, \ldots, x_n) \coloneqq x_i.$$

This gives us functions *out of* a product. We also wish to be able to define functions *into* a product, via tuples of functions: if we have a set Y and n functions $f_1 \colon Y \to X_1$, ..., $f_n \colon Y \to X_n$, then we can form a new function $Y \to X_1 \times \cdots \times X_n$, namely:

$$Y \xrightarrow{\langle f_1, \ldots, f_n \rangle} X_1 \times \cdots \times X_n \quad \text{via} \quad \langle f_1, \ldots, f_n \rangle(y) \coloneqq (f_1(y), \ldots, f_n(y)).$$

There is an obvious result about projecting after tupling of functions:

$$\pi_i \circ \langle f_1, \ldots, f_n \rangle = f_i. \tag{1.1}$$

This is an equality of functions. It can be proven easily by applying both sides to an arbitrary element $y \in Y$.

There are some more 'obvious' equations about tupling of functions:

$$\langle f_1, \ldots, f_n \rangle \circ g = \langle f_1 \circ g, \ldots, f_n \circ g \rangle \qquad \langle \pi_1, \ldots, \pi_n \rangle = id, \tag{1.2}$$

where $g \colon Z \to Y$ is an arbitrary function. In the last equation, id is the identity function on the product $X_1 \times \cdots \times X_n$.

In a Cartesian product we place sets 'in parallel'. We can also place functions

between them in parallel. Suppose we have n functions $f_i \colon X_i \to Y_i$. Then we can form the parallel composition:

$$X_1 \times \cdots \times X_n \xrightarrow{f_1 \times \cdots \times f_n} Y_1 \times \cdots \times Y_n$$

via:

$$f_1 \times \cdots \times f_n = \langle f_1 \circ \pi_1, \ldots, f_n \circ \pi_n \rangle$$

so that:

$$(f_1 \times \cdots \times f_n)(x_1, \ldots, x_n) = (f_1(x_1), \ldots, f_n(x_n)).$$

The latter formulation clearly shows how the functions f_i are applied in parallel to the elements x_i.

We overload the product symbol $\times$, since we use it both for sets and for functions. This may be a bit confusing at first, but it is in fact quite convenient.

1.1.2 Powers and exponents

Let X be an arbitrary set. A *power* of X is an n-product of X's, for some n. We write the n-th power of X as X^n, in:

$$X^n \coloneqq \underbrace{X \times \cdots \times X}_{n \text{ times}} = \{(x_1, \ldots, x_n) \mid x_i \in X \text{ for each } i\}.$$

As special cases we have $X^1 = X$ and $X^0 = 1$, where $1 = \{()\}$ is the singleton set with the empty tuple $()$ as sole inhabitant. Since powers are special cases of Cartesian products, they come with projection functions $\pi_i \colon X^n \to X$ and tuple functions $\langle f_1, \ldots, f_n \rangle \colon Y \to X^n$ for n functions $f_i \colon Y \to X$. Finally, for a function $f \colon X \to Y$ we write $f^n \colon X^n \to Y^n$ for the obvious n-fold parallelisation of f.

More generally, for two sets X, Y we shall occasionally write:

$$X^Y \coloneqq \{\, \text{functions } f \colon Y \to X \,\}.$$

This new set X^Y is sometimes called the *function space* or the *exponent* of X and Y. Notice that this exponent notation is consistent with the above one for powers, since functions $n \to X$ can be identified with n-tuples of elements in X.

These exponents X^Y are related to products in an elementary and useful way, namely via a bijective correspondence:

$$\frac{Z \times Y \xrightarrow{f} X}{Z \xrightarrow[g]{} X^Y} \tag{1.3}$$

This means that for a function $f\colon Z \times Y \to X$ there is a corresponding function $\overline{f}\colon Z \to X^Y$, and vice-versa, for $g\colon Z \to X^Y$ there is a function $\overline{g}\colon Z \times Y \to X$, in such a way that $\overline{\overline{f}} = f$ and $\overline{\overline{g}} = g$. It is not hard to see that we can take $\overline{f}(z) \in X^Y$ to be the function $\overline{f}(z)(y) = f(z, y)$, for $z \in Z$ and $y \in Y$. Similarly, we use $\overline{g}(z, y) = g(z)(y)$.

The correspondence (1.3) is characteristic for so-called Cartesian closed categories.

Exercises

1.1.1 Check what a tuple function $\langle \pi_2, \pi_3, \pi_6 \rangle$ does on a product set $X_1 \times \cdots \times X_8$. What is the codomain of this function?

1.1.2 Check that, in general, the tuple function $\langle f_1, \ldots, f_n \rangle$ is the *unique* function $h\colon Y \to X_1 \times \cdots \times X_n$ with $\pi_i \circ h = f_i$ for each i.

1.1.3 Prove, using Equations (1.1) and (1.2) for tuples and projections, that:

$$(g_1 \times \cdots \times g_n) \circ (f_1 \times \cdots \times f_n) = (g_1 \circ f_1) \times \cdots \times (g_n \circ f_n).$$

1.1.4 Check that for each set X there is a unique function $X \to 1$. Because of this property the set 1 is sometimes called 'final' or 'terminal'. The unique function is often denoted by !.

Check also that a function $1 \to X$ corresponds to an element of X.

1.1.5 Define functions in both directions, using tuples and projections, that yield isomorphisms:

$$X \times Y \cong Y \times X \qquad 1 \times X \cong X \qquad X \times (Y \times Z) \cong (X \times Y) \times Z.$$

Try to use Equations (1.1) and (1.2) to prove these isomorphisms, without reasoning with elements.

1.1.6 Similarly, show that exponents satisfy:

$$X^1 \cong X \qquad 1^Y \cong 1 \qquad (X \times Y)^Z \cong X^Z \times Y^Z \qquad X^{Y \times Z} \cong (X^Y)^Z.$$

1.1.7 For $K \in \mathbb{N}$ and sets X, Y define:

$$X^K \times Y^K \xrightarrow{\mathit{zip}[K]} (X \times Y)^K$$

by:

$$\mathit{zip}[K]((x_1, \ldots, x_K), (y_1, \ldots, y_K)) \coloneqq ((x_1, y_1), \ldots, (x_K, y_K)).$$

Show that *zip* is an isomorphism, with inverse function $\mathit{unzip}[K] \coloneqq \langle (\pi_1)^K, (\pi_2)^K \rangle$.

1.2 Lists

The datatype of (finite) lists of elements from a given set is well-known in computer science, especially in functional programming. This section collects some basic constructions and properties, especially about the close relationship between lists and monoids.

For an arbitrary set X we write $\mathcal{L}(X)$ for the set of all finite lists $[x_1, \ldots, x_n]$ of elements $x_i \in X$, for arbitrary $n \in \mathbb{N}$. Notice that we use square brackets $[-]$ for lists, to distinguish them from tuples, which are typically written with round brackets $(-)$.

Thus, the set of lists over X can be defined as a union of all powers of X, as in:

$$\mathcal{L}(X) \coloneqq \bigcup_{n \in \mathbb{N}} X^n.$$

When the elements of X are letters of an alphabet, then $\mathcal{L}(X)$ is the set of words — the language — over this alphabet. The set $\mathcal{L}(X)$ is alternatively written as $X^\star$, and called the Kleene star of X.

We zoom in on some trivial cases. One has $\mathcal{L}(0) \cong 1$, since one can only form the empty word over the empty alphabet $0 = \emptyset$. If the alphabet contains only one letter, a word consists of a finite number of occurrences of this single letter. Thus: $\mathcal{L}(1) \cong \mathbb{N}$.

We consider lists as an instance of what we call a collection data type, since $\mathcal{L}(X)$ collects elements of X in a certain manner. What distinguishes lists from other collection types is that elements may occur multiple times, and that the order of occurrence matters. The three lists $[a, b, a]$, $[a, a, b]$, and $[a, b]$ differ. As mentioned in the introduction to this chapter, within a *subset* orders and multiplicities do not matter, see Section 1.3; and in a *multiset* the order of elements does not matter, but multiplicities do matter, see Section 1.4.

Let $f \colon X \to Y$ be an arbitrary function. It can be used to map lists over X into lists over Y by applying f element-wise. This is what functional programmers call map-list. Here we like overloading, so we write $\mathcal{L}(f) \colon \mathcal{L}(X) \to \mathcal{L}(Y)$ for this function, defined as:

$$\mathcal{L}(f)([x_1, \ldots, x_n]) \coloneqq [f(x_1), \ldots, f(x_n)].$$

Thus, $\mathcal{L}$ is an operation that not only sends sets to sets, but also functions to functions. It does so in such a way that identity maps and compositions are preserved:

$$\mathcal{L}(id) = id \qquad \mathcal{L}(g \circ f) = \mathcal{L}(g) \circ \mathcal{L}(f).$$

We shall say: the operation $\mathcal{L}$ is *functorial*, or simply $\mathcal{L}$ is a *functor*.

Functoriality can be used to define the *marginal* of a list on a product set, via $\mathcal{L}(\pi_i)$, where π_i is a projection map. For instance, let $\ell \in \mathcal{L}(X \times Y)$ be of the form $\ell = [(x_1, y_1), \ldots, (x_n, y_n)]$. The first marginal $\mathcal{L}(\pi_1)(\ell) \in \mathcal{L}(X)$ is then computed as:

$$\begin{aligned} \mathcal{L}(\pi_1)(\ell) &= \mathcal{L}(\pi_1)([(x_1, y_1), \ldots, (x_n, y_n)]) \\ &= [\pi_1(x_1, y_1), \ldots, \pi_1(x_n, y_n)] \\ &= [x_1, \ldots, x_n]. \end{aligned}$$

1.2.1 Monoids

A monoid is a very basic mathematical structure. For convenience we define it explicitly.

Definition 1.2.1. A *monoid* consists of a set M with a binary operation $M \times M \to M$, written for instance as infix $+$, together with an identity element, say written as $0 \in M$. The binary operation $+$ is associative and has 0 as identity on both sides. That is, for all $a, b, c \in M$,

$$a + (b + c) = (a + b) + c \qquad \text{and} \qquad 0 + a = a = a + 0.$$

The monoid is called *commutative* if $a + b = b + a$, for all $a, b \in M$. It is called *idempotent* if $a + a = a$ for all $a \in M$.

Let $(M, 0, +)$ and $(N, 1, \cdot)$ be two monoids. A function $f \colon M \to N$ is called a *homomorphism of monoids* if f preserves the unit and binary operation, in the sense that:

$$f(0) = 1 \qquad \text{and} \qquad f(a + b) = f(a) \cdot f(b), \quad \text{for all } a, b \in M.$$

For brevity we also say that such an f is a *map of monoids*, or simply a *monoid map*.

The natural numbers $\mathbb{N}$ with addition form a commutative monoid $(\mathbb{N}, 0, +)$. But also with multiplication they form a commutative monoid $(\mathbb{N}, 1, \cdot)$. The function $f(n) = 2^n$ is a homomorphism of monoids $f \colon (\mathbb{N}, 0, +) \to (\mathbb{N}, 1, \cdot)$.

Various forms of collection types form monoids, with 'union' as binary operation. We start with lists, in the next result. The proof is left as (an easy) exercise to the reader.

Lemma 1.2.2. *1 For each set X, the set $\mathcal{L}(X)$ of lists over X is a monoid, with the empty list $[] \in \mathcal{L}(X)$ as identity element, and with concatenation $+\!+ \colon \mathcal{L}(X) \times \mathcal{L}(X) \to \mathcal{L}(X)$ as binary operation:*

$$[x_1, \ldots, x_n] +\!+ [y_1, \ldots, y_m] \coloneqq [x_1, \ldots, x_n, y_1, \ldots, y_m].$$

This monoid $(\mathcal{L}(X), [], +\!\!+)$ *is neither commutative nor idempotent.*

2 *For each function* $f\colon X \to Y$ *the associated map* $\mathcal{L}(f)\colon \mathcal{L}(X) \to \mathcal{L}(Y)$ *is a homomorphism of monoids.* ❑

Thus, lists are monoids via concatenation. But there is more to say: lists are *free* monoids. We shall occasionally make use of this basic property and so we like to make it explicit. We shall encounter similar freeness properties for other collection types.

Each element $x \in X$ yields a singleton list $\mathit{unit}(x) \coloneqq [x] \in \mathcal{L}(X)$. The resulting function $\mathit{unit}\colon X \to \mathcal{L}(X)$ plays a special role, see also the next subsection.

Proposition 1.2.3. *Let* X *be an arbitrary set and let* $(M, 0, +)$ *be an arbitrary monoid, with a function* $f\colon X \to M$*. Then there is a unique homomorphism of monoids* $\overline{f}\colon (\mathcal{L}(X), [], +\!\!+) \to (M, 0, +)$ *with* $\overline{f} \circ \mathit{unit} = f$*.*

The homomorphism $\overline{f}$ *is called the* free extension *of* f*. Its freeness can be expressed via a diagram, as below, where the vertical arrow is dashed, to indicate uniqueness.*

$$\begin{array}{ccc} X & \xrightarrow{\ \mathit{unit}\ } & \mathcal{L}(X) \\ & \searrow^{f} & \big\downarrow \overline{f},\ \textit{homomorphism} \\ & & M \end{array} \tag{1.4}$$

Proof. Since $\overline{f}$ preserves the identity element and satisfies $\overline{f} \circ \mathit{unit} = f$ it is determined on empty and singleton lists as:

$$\overline{f}([]) = 0 \qquad \text{and} \qquad \overline{f}([x]) = f(x).$$

Further, on an list $[x_1, \ldots, x_n]$ of length $n \geq 2$ we necessarily have:

$$\begin{aligned} \overline{f}\big([x_1, \ldots, x_n]\big) &= \overline{f}([x_1] +\!\!+ \cdots +\!\!+ [x_n]) \\ &= \overline{f}([x_1]) + \cdots + \overline{f}([x_n]) \\ &= f(x_1) + \cdots + f(x_n). \end{aligned}$$

Thus, there is only one way to define $\overline{f}$. By construction, this $\overline{f}\colon \mathcal{L}(X) \to M$ is a homomorphism of monoids. ❑

The exercise below illustrate this result. For future use we introduce monoid actions and their homomorphisms.

Definition 1.2.4. Let $(M, 0, +)$ be a monoid.

1 An *action* of the monoid M on a set X is a function $\alpha\colon M \times X \to X$ satisfying:

$$\alpha(0, x) = x \qquad \text{and} \qquad \alpha(a + b, x) = \alpha(a, \alpha(b, x)),$$

for all $a, b \in M$ and $x \in X$.

2 A *homomorphism* or *map* of monoid actions from $(M \times X \xrightarrow{\alpha} X)$ to $(M \times Y \xrightarrow{\beta} Y)$ is a function $f\colon X \to Y$ satisfing:

$$f(\alpha(a, x)) = \beta(a, f(x)) \qquad \text{for all } a \in M, x \in X.$$

This equation corresponds to commutation of the following diagram.

$$\begin{array}{ccc} M \times X & \xrightarrow{id \times f} & M \times Y \\ {\scriptstyle\alpha}\downarrow & & \downarrow{\scriptstyle\beta} \\ X & \xrightarrow{f} & Y \end{array}$$

Monoid actions are quite common in mathematics. For instance, scalar multiplication of a vector space forms an action. Also, as we shall see, probabilistic updating can be described via monoid actions. The action map $\alpha\colon M \times X \to X$ can be understood intuitively as pushing the elements in X forward with a quantity from M. It then makes sense that the zero-push is the identity, and that a sum-push is the composition of two individual pushes.

1.2.2 Unit and flatten for lists

We proceed to describe more elementary structure for lists, in terms of special 'unit' and 'flatten' functions. In subsequent sections we shall see that this same structure exists for other collection types, like powerset, multiset and distribution. This unit and flatten structure will turn out to be essential for sequential composition. At the end of this chapter we will see that it is characteristic for what is called a 'monad'.

We have already seen the singleton-list function $\mathit{unit}\colon X \to \mathcal{L}(X)$, given $\mathit{unit}(x) \coloneqq [x]$. There is also a 'flatten' function which turns a list of lists into a list by removing inner brackets. This function is written as $\mathit{flat}\colon \mathcal{L}(\mathcal{L}(X)) \to \mathcal{L}(X)$. It is defined as:

$$\mathit{flat}([[x_{11}, \ldots, x_{1n_1}], \ldots, [x_{k1}, \ldots, x_{kn_k}]]) \coloneqq [x_{11}, \ldots, x_{1n_1}, \ldots, x_{k1}, \ldots, x_{kn_k}].$$

The next result contains some basic properties about unit and flatten. These properties will first be formulated in terms of equations, and then, alternatively as commuting diagrams. The latter style is preferred in this book.

Lemma 1.2.5. *1 For each function $f\colon X \to Y$ one has:*

$$\mathit{unit} \circ f = \mathcal{L}(f) \circ \mathit{unit} \qquad \text{and} \qquad \mathit{flat} \circ \mathcal{L}(\mathcal{L}(f)) = \mathcal{L}(f) \circ \mathit{flat}.$$

Equivalently, the following two diagrams commute.

$$\begin{array}{ccc} X & \xrightarrow{unit} & \mathcal{L}(X) \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle \mathcal{L}(f)} \\ Y & \xrightarrow[unit]{} & \mathcal{L}(Y) \end{array} \qquad \begin{array}{ccc} \mathcal{L}(\mathcal{L}(X)) & \xrightarrow{flat} & \mathcal{L}(X) \\ {\scriptstyle \mathcal{L}(\mathcal{L}(f))}\downarrow & & \downarrow{\scriptstyle \mathcal{L}(f)} \\ \mathcal{L}(\mathcal{L}(Y)) & \xrightarrow[flat]{} & \mathcal{L}(Y) \end{array}$$

2 *One further has:*

$$flat \circ unit \;=\; id \;=\; flat \circ \mathcal{L}(unit) \qquad flat \circ flat \;=\; flat \circ \mathcal{L}(flat).$$

These two equations can equivalently be expressed via commutation of:

$$\begin{array}{ccccc} \mathcal{L}(X) & \xrightarrow{unit} & \mathcal{L}(\mathcal{L}(X)) & \xleftarrow{\mathcal{L}(unit)} & \mathcal{L}(X) \\ & \searrow & \downarrow{\scriptstyle flat} & \swarrow & \\ & & \mathcal{L}(X) & & \end{array} \qquad \begin{array}{ccc} \mathcal{L}(\mathcal{L}(\mathcal{L}(X))) & \xrightarrow{flat} & \mathcal{L}(\mathcal{L}(X)) \\ {\scriptstyle \mathcal{L}(flat)}\downarrow & & \downarrow{\scriptstyle flat} \\ \mathcal{L}(\mathcal{L}(X)) & \xrightarrow[flat]{} & \mathcal{L}(X) \end{array}$$

Proof. We shall do the first cases of each item, leaving the second cases to the interested reader. First, for $f\colon X \to Y$ and $x \in X$ one has:

$$\begin{aligned} (\mathcal{L}(f) \circ unit)(x) \;=\; \mathcal{L}(f)(unit(x)) \;&=\; \mathcal{L}(f)([x]) \\ &=\; [f(x)] \;=\; unit(f(x)) \;=\; (unit \circ f)(x). \end{aligned}$$

Next, for the second item we take an arbitrary list $[x_1, \ldots, x_n] \in \mathcal{L}(X)$. Then:

$$\begin{aligned} (flat \circ unit)([x_1, \ldots, x_n]) \;&=\; flat([[x_1, \ldots, x_n]]) \\ &=\; [x_1, \ldots, x_n] \\ (flat \circ \mathcal{L}(unit))([x_1, \ldots, x_n]) \;&=\; flat([unit(x_1), \ldots, unit(x_n)]) \\ &=\; flat([[x_1], \ldots, [x_n]]) \\ &=\; [x_1, \ldots, x_n]. \end{aligned}$$

❑

The equations in item 1 of this lemma are so-called naturality equations. They express that *unit* and *flat* work uniformly, independent of the set X involved. The equations in item 2 show that $\mathcal{L}$ is a monad, see Section 1.9 for more information.

The next result connects monoids with the unit and flatten maps.

Proposition 1.2.6. *Let X be an arbitrary set.*

1 *To give a monoid structure $(u, +)$ on X is the same as giving an $\mathcal{L}$-*algebra*, that is, a map $\alpha\colon \mathcal{L}(X) \to X$ satisfying $\alpha \circ unit = id$ and $\alpha \circ flat = \alpha \circ$*

$\mathcal{L}(\alpha)$*, as in:*

$$\begin{array}{ccc} X & \xrightarrow{\mathit{unit}} & \mathcal{L}(X) \\ & \searrow^{\mathit{id}} & \downarrow{\scriptstyle \alpha} \\ & & X \end{array} \qquad\qquad \begin{array}{ccc} \mathcal{L}(\mathcal{L}(X)) & \xrightarrow{\mathcal{L}(\alpha)} & \mathcal{L}(X) \\ {\scriptstyle \mathit{flat}}\downarrow & & \downarrow{\scriptstyle \alpha} \\ \mathcal{L}(X) & \xrightarrow{\alpha} & X \end{array} \tag{1.5}$$

2 *Let* $(M_1, u_1, +_1)$ *and* $(M_2, u_2, +_2)$ *be two monoids, with corresponding* $\mathcal{L}$*-algebras* $\alpha_1 \colon \mathcal{L}(M_1) \to M_1$ *and* $\alpha_2 \colon \mathcal{L}(M_2) \to M_2$*. A function* $f \colon M_1 \to M_2$ *is then a homomorphism of monoids if and only if the diagram*

$$\begin{array}{ccc} \mathcal{L}(M_1) & \xrightarrow{\mathcal{L}(f)} & \mathcal{L}(M_2) \\ {\scriptstyle \alpha_1}\downarrow & & \downarrow{\scriptstyle \alpha_2} \\ M_1 & \xrightarrow{f} & M_2. \end{array} \tag{1.6}$$

commutes.

This result says that instead of giving a binary operation + with an identity element u we can give a single operation α that works on all sequences of elements. This is not so surprising, since we can apply the sum multiple times. The more interesting part is that the monoid equations can be captured uniformly by the diagrams/equations (1.5). We shall see that same diagrams also work for other types of monoids (and collection types).

Proof. 1 If $(X, u, +)$ is a monoid, we can define $\alpha \colon \mathcal{L}(X) \to X$ in one go as $\alpha([x_1, \ldots, x_n]) \coloneqq x_1 + \cdots + x_n$. The latter sum equals the identity element u when $n = 0$. Notice that the bracketing of the elements in the expression $x_1 + \cdots + x_n$ does not matter, since a monoid is associative. The order does matter, since we do not assume that the monoid is commutative. It is easy to check the equations (1.5). From a more abstract perspective we define $\alpha \colon \mathcal{L}(X) \to X$ via freeness (1.4).

In the other direction, assume an $\mathcal{L}$-algebra $\alpha \colon \mathcal{L}(X) \to X$. We then define an identity element $u \coloneqq \alpha([]) \in X$ and the sum of $x, y \in X$ as $x + y \coloneqq \alpha([x, y]) \in X$. We have to check that u is identity for + and that + is associative. This requires some fiddling with the equations (1.5):

$$\begin{aligned} x + u \;=\; \alpha([x, \alpha([])]) &\overset{(1.5)}{=} \alpha([\alpha(\mathit{unit}(x)), \alpha([])]) \\ &= \alpha\Big(\mathcal{L}(\alpha)([\,[x], []\,])\Big) \\ &\overset{(1.5)}{=} \alpha\Big(\mathit{flat}([\,[x], []\,])\Big) \;=\; \alpha([x]) \overset{(1.5)}{=} x. \end{aligned}$$

Similarly one shows $u + y = y$. Next, associativity of + is obtained in a

similar manner:

$$\begin{aligned} x + (y + z) \;=\; \alpha([x, \alpha([y, z])]) &\overset{(1.5)}{=} \alpha([\alpha(\mathit{unit}(x)), \alpha([y, z])]) \\ &= \alpha\Big(\mathcal{L}(\alpha)([\,[x], [y, z]\,])\Big) \\ &\overset{(1.5)}{=} \alpha\Big(\mathit{flat}([\,[x], [y, z]\,])\Big) \\ &= \alpha\Big(\mathit{flat}([\,[x, y], [z]\,])\Big) \\ &\overset{(1.5)}{=} \alpha\Big(\mathcal{L}(\alpha)([\,[x, y], [z]\,])\Big) \\ &= \alpha([\alpha([x, y]), \alpha(\mathit{unit}(z))]) \\ &\overset{(1.5)}{=} \alpha([\alpha([x, y]), z]) \;=\; (x + y) + z. \end{aligned}$$

2 Now let $f\colon M_1 \to M_2$ be a homomorphism of monoids. Diagram (1.6) then commutes:

$$\begin{aligned} (\alpha_2 \circ \mathcal{L}(f))([x_1, \ldots, x_n]) &= \alpha_2([f(x_1), \ldots, f(x_n)]) \\ &= f(x_1) + \cdots + f(x_n) \\ &= f(x_1 + \cdots + x_n) \quad \text{since } f \text{ is a homomorphism} \\ &= (f \circ \alpha_1)([x_1, \ldots, x_n]). \end{aligned}$$

In the other direction, if (1.6) commutes for a function $f\colon M_1 \to M_2$, then f is a homomorphism of monoids, since:

$$f(u_1) \;=\; f(\alpha_1([])) \overset{(1.6)}{=} \alpha_2(\mathcal{L}(f)([])) \;=\; \alpha_2([]) \;=\; u_2.$$

Similarly one checks that sums are preserved:

$$\begin{aligned} f(x +_1 y) \;=\; f(\alpha_1([x, y])) &\overset{(1.6)}{=} \alpha_2(\mathcal{L}(f)([x, y])) \\ &= \alpha_2([f(x), f(y)]) \;=\; f(x) +_2 f(y). \end{aligned}$$ ❑

We see that the algebraic structure $(M, u, +)$ on the set M is expressed as an algebra $\alpha\colon \mathcal{L}(M) \to M$, namely a certain map to M. This will be a recurring theme in the coming sections.

1.2.3 List combinatorics

Combinatorics is a subarea of mathematics focused on advanced forms of counting. It is relevant for probability theory, since frequencies of occurrences play an important role. We give a first taste of this, using lists.

We shall use the length $\|\ell\| \in \mathbb{N}$ of a list ℓ, see Exercise 1.2.3 for more details, and also the sum and product of a list of natural numbers, defined as:

$$\begin{aligned} \mathit{sum}\Big([n_1, \ldots, n_k]\Big) &:= n_1 + \cdots + n_k \;=\; \textstyle\sum_i n_i \\ \mathit{prod}\Big([n_1, \ldots, n_k]\Big) &:= n_1 \cdot \ldots \cdot n_k \;=\; \textstyle\prod_i n_i. \end{aligned}$$

See also Exercise 1.2.6.

We restrict ourselves to the subset $\mathbb{N}_{>0} = \{n \in \mathbb{N} \mid n > 0\}$ of positive natural numbers. Clearly, we obtain restrictions $sum, prod \colon \mathcal{L}(\mathbb{N}_{>0}) \to \mathbb{N}_{>0}$ of the above sum and product functions.

Now fix $N \in \mathbb{N}_{>0}$. We are interested in lists $\ell \in \mathcal{L}(\mathbb{N}_{>0})$ with $sum(\ell) = N$. These lists are in the inverse image:

$$sum^{-1}(N) \coloneqq \{\ell \in \mathcal{L}(\mathbb{N}_{>0}) \mid sum(\ell) = N\}.$$

For instance, for $N = 4$ this inverse image contains the eight lists:

$$[1,1,1,1],\ [1,1,2],\ [1,2,1],\ [2,1,1],\ [2,2],\ [1,3],\ [3,1],\ [4]. \tag{1.7}$$

We can interpret the situation as follows. Suppose we have coins with value $n \in \mathbb{N}_{>0}$, for each n. Then we can ask, for an amount N: how many (ordered) ways are there to lay out the amount N in coins? For $N = 4$ the different layouts are given above. Other interpretations are possible: one can also think of the sequences (1.7) as partitions of the numbers $\{1, 2, 3, 4\}$.

Here is a first, easy counting result.

Lemma 1.2.7. *For $N \in \mathbb{N}_{>0}$, the subset $sum^{-1}(N) \subseteq \mathcal{L}(\mathbb{N}_{>0})$ has 2^{N-1} elements.*

Proof. We use induction on N, starting with $N = 1$. Obviously, only the list $[1]$ sums to 1, and indeed $2^{N-1} = 2^0 = 1$.

For the induction step we use the familiar fact that for $K \in \mathbb{N}$,

$$\sum_{0 \leq k \leq K} \frac{1}{2^k} = \frac{2^{K+1} - 1}{2^K}. \tag{$*$}$$

This can be shown easily by induction on K.

We anticipate the notation $|A|$ for the number of elements of a finite set A,

see Definition 1.3.6. Then, for $N > 1$,

$$\begin{aligned}
\left|sum^{-1}(N)\right| &= \left|\{[N]\} \cup \{\ell +\!\!+ [n] \,\middle|\, 1 \leq n \leq N-1 \text{ and } \ell \in sum^{-1}(N-n)\}\right| \\
&= 1 + \sum_{1 \leq n \leq N-1} \left|sum^{-1}(N-n)\right| \\
&\overset{\text{(IH)}}{=} 1 + \sum_{1 \leq n \leq N-1} 2^{N-n-1} \\
&= 1 + 2^{N-1} \cdot \left(\sum_{0 \leq n \leq N-1} \frac{1}{2^n} \;-\; 1 \right) \\
&\overset{(*)}{=} 1 + 2^{N-1} \cdot \left(\frac{2^N - 1}{2^{N-1}} \;-\; 1 \right) \\
&= 1 + 2^N - 1 - 2^{N-1} \\
&= 2^{N-1}.
\end{aligned}$$

❑

Here is an elementary fact about coin lists. It has a definite probabilistic flavour, since it involves what we later call a convex sum of probabilities $r_i \in [0, 1]$ with $\sum_i r_i = 1$. The proof of this result is postponed. It involves rather complex probability distributions about coins, see Corollary 3.9.13. We are not aware of an elementary proof.

Theorem 1.2.8. *For each $N \in \mathbb{N}_{>0}$,*

$$\sum_{\ell \in sum^{-1}(N)} \frac{1}{\|\ell\|! \cdot prod(\ell)} \;=\; 1. \tag{1.8}$$

At this stage we only give an example, for $N = 4$, using the corresponding lists ℓ in (1.7). The associated sum (1.8) is illustrated below.

$[1,1,1,1]$	$[1,1,2]$	$[1,2,1]$	$[2,1,1]$	$[2,2]$	$[1,3]$	$[3,1]$	$[4]$
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
$\frac{1}{4!\cdot 1}$	$\frac{1}{3!\cdot 2}$	$\frac{1}{3!\cdot 2}$	$\frac{1}{3!\cdot 2}$	$\frac{1}{2!\cdot 4}$	$\frac{1}{2!\cdot 3}$	$\frac{1}{2!\cdot 3}$	$\frac{1}{1!\cdot 4}$
$\|\|$	$\|\|$	$\|\|$	$\|\|$	$\|\|$	$\|\|$	$\|\|$	$\|\|$
$\frac{1}{24}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{4}$

with sum: 1

Obviously, elements in a lists are ordered. Thus, in (1.7) we distinghuish between coin layouts $[1, 1, 2]$, $[1, 2, 1]$ and $[2, 1, 1]$. However, when we are commonly discussing which coins add up to 4 we do not take the order into account, for instance in saying that we use two coins of value 1 and one coin of 2, without caring about the order. In doing so, we are not using lists as collection type, but multisets — in which the order of elements does not matter.

These multisets form an important alternative collection type; they are discussed from Section 1.4 onwards.

Exercises

1.2.1 Let $X = \{a, b, c\}$ and $Y = \{u, v\}$ be sets with a function $f\colon X \to Y$ given by $f(a) = u = f(c)$ and $f(b) = v$. Write $\ell_1 = [c, a, b, a]$ and $\ell_2 = [b, c, c, c]$. Compute consecutively:

- $\ell_1 +\!\!+ \ell_2$
- $\ell_2 +\!\!+ \ell_1$
- $\ell_1 +\!\!+ \ell_1$
- $\ell_1 +\!\!+ (\ell_2 +\!\!+ \ell_1)$
- $(\ell_1 +\!\!+ \ell_2) +\!\!+ \ell_1$
- $\mathcal{L}(f)(\ell_1)$
- $\mathcal{L}(f)(\ell_2)$
- $\mathcal{L}(f)(\ell_1) +\!\!+ \mathcal{L}(f)(\ell_2)$
- $\mathcal{L}(f)(\ell_1 +\!\!+ \ell_2)$.

1.2.2 We write log for the logarithm function with some base $b > 0$, so that $\log(x) = y$ iff $x = b^y$. Verify that the logarithm function log is a map of monoids:

$$(\mathbb{R}_{>0}, 1, \cdot) \xrightarrow{\log} (\mathbb{R}, 0, +).$$

Often the log function is used to simplify a computation, by turning multiplications into additions. Then one uses that log is precisely this homomorphism of monoids. (An additonal useful property is that log is monotone: it preserves the order.)

1.2.3 Define a length function $\| - \|\colon \mathcal{L}(X) \to \mathbb{N}$ on lists via the freeneess property of Proposition 1.2.3.

1 Describe $\|\ell\|$ for $\ell \in \mathcal{L}(X)$ explicitly.
2 Elaborate what it means that $\| - \|$ is a homomorphism of monoids.
3 Write ! for the unique function $X \to 1$ and check that $\| - \|$ is $\mathcal{L}(!)$. Notice that the previous item can then be seen as an instance of Lemma 1.2.2 (2).

1.2.4 1 Check that the list-flatten operation $\mathcal{L}(\mathcal{L}(X)) \to \mathcal{L}(X)$ can be described in terms of concatenation $+\!\!+$ as:

$$\mathit{flat}([\ell_1, \ldots, \ell_n]) = \ell_1 +\!\!+ \cdots +\!\!+ \ell_n.$$

2 Now consider the correspondence of Proposition 1.2.6 (1). Conclude that the algebra $\alpha\colon \mathcal{L}(\mathcal{L}(X)) \to \mathcal{L}(X)$ associated with the monoid $(\mathcal{L}(X), [], ++)$ from Lemma 1.2.2 (1) is *flat*.

1.2.5 Check Equation (1.8) yourself for $N = 5$.

1.2.6 Consider the set $\mathbb{N}$ of natural numbers with its *additive* monoid structure $(0, +)$ and also with its *multiplicative* monoid structure $(1, \cdot)$. Apply freeness from Proposition 1.2.3 with these two structures to define two monoid homomorphisms:

$$\mathcal{L}(\mathbb{N}) \;\underset{prod}{\overset{sum}{\rightrightarrows}}\; \mathbb{N}$$

1 Describe these maps explicitly on a sequence $[n_1, \ldots, n_k]$ of natural numbers n_i.

2 Make explicit what it means that they preserve the monoid structure.

3 Prove that for an arbitrary set X, the list-length function $\| - \|$ from Exercise 1.2.3 satisfies:

$$sum \circ \mathcal{L}(\| - \|) = \| - \| \circ flat.$$

In other words, the following diagram commutes.

$$\begin{array}{ccc} \mathcal{L}(\mathcal{L}(X)) & \xrightarrow{\mathcal{L}(\|-\|)} & \mathcal{L}(\mathbb{N}) \\ {\scriptstyle flat}\downarrow & & \downarrow{\scriptstyle sum} \\ \mathcal{L}(X) & \xrightarrow{\|-\|} & \mathbb{N} \end{array}$$

A fancy way to prove that length is such an algebra homomorphism is to use the uniqueness in Proposition 1.2.3.

1.3 Subsets

The next collection type that will be studied is powerset. The symbol $\mathcal{P}$ is commonly used for the powerset operator. We will see that there are many similarities with lists $\mathcal{L}$ from the previous section. We again pay much attention to monoid structures.

For an arbitrary set X we write $\mathcal{P}(X)$ for the set of *all* subsets of X, and $\mathcal{P}_{fin}(X)$ for the set of *finite* subsets. Thus:

$$\mathcal{P}(X) := \{U \mid U \subseteq X\} \qquad \text{and} \qquad \mathcal{P}_{fin}(X) := \{U \in \mathcal{P}(X) \mid U \text{ is finite}\}.$$

If X is a finite set itself, there is no difference between $\mathcal{P}(X)$ and $\mathcal{P}_{fin}(X)$. In the sequel we shall speak mostly about $\mathcal{P}$, but basically all properties of interest hold for $\mathcal{P}_{fin}$ as well.

First of all, $\mathcal{P}$ is a functor: it works both on sets and on functions. Given a function $f\colon X \to Y$ we can define a new function $\mathcal{P}(f)\colon \mathcal{P}(X) \to \mathcal{P}(Y)$ by taking the image of f on a subset. Explicitly, for $U \subseteq X$,

$$\mathcal{P}(f)(U) = \{f(x) \mid x \in U\}.$$

The right-hand side is clearly a subset of Y, and thus an element of $\mathcal{P}(Y)$. We have two equalities:

$$\mathcal{P}(id) = id \qquad \mathcal{P}(g \circ f) = \mathcal{P}(g) \circ \mathcal{P}(f).$$

We can use functoriality for marginalisation: for a subset (relation) $R \subseteq X \times Y$ on a product set we get its first marginal $\mathcal{P}(\pi_1)(R) \in \mathcal{P}(X)$ as the subset:

$$\mathcal{P}(\pi_1)(R) = \{\pi_1(z) \mid z \in R\} = \{\pi_1(x, y) \mid (x, y) \in R\} = \{x \mid \exists y.\, (x, y) \in R\}.$$

The next topic is the monoid structure on powersets. The first result is an analogue of Lemma 1.2.2 and its proof is left to the reader.

Lemma 1.3.1. *1 For each set X, the powerset $\mathcal{P}(X)$ is a commutative and idempotent monoid, with empty subset $\emptyset \in \mathcal{P}(X)$ as identity element and union $\cup$ of subsets of X as binary operation.*

2 Each $\mathcal{P}(f)\colon \mathcal{P}(X) \to \mathcal{P}(Y)$ is a map of monoids, for $f\colon X \to Y$. ❑

Next we define unit and flatten maps for subsets, much like for lists. The function $unit\colon X \to \mathcal{P}(X)$ sends an element to a singleton subset: $unit(x) := \{x\}$. The flatten function $flat\colon \mathcal{P}(\mathcal{P}(X)) \to \mathcal{P}(X)$ is given by union: for $A \subseteq \mathcal{P}(X)$,

$$flat(A) := \bigcup A = \{x \in X \mid \exists U \in A.\, x \in U\}.$$

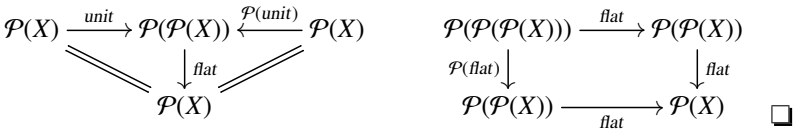
We mention, without proof, the following analogue of Lemma 1.2.5.

Lemma 1.3.2. *1 For each function $f\colon X \to Y$ the 'naturality' diagrams*

$$\begin{array}{ccc} X & \xrightarrow{unit} & \mathcal{P}(X) \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle \mathcal{P}(f)} \\ Y & \xrightarrow[unit]{} & \mathcal{P}(Y) \end{array} \qquad\qquad \begin{array}{ccc} \mathcal{P}(\mathcal{P}(X)) & \xrightarrow{flat} & \mathcal{P}(X) \\ {\scriptstyle \mathcal{P}(\mathcal{P}(f))}\downarrow & & \downarrow{\scriptstyle \mathcal{P}(f)} \\ \mathcal{P}(\mathcal{P}(Y)) & \xrightarrow[flat]{} & \mathcal{P}(Y) \end{array}$$

commute.

2 Additionaly, the 'monad' diagrams below commute.



❑

1.3.1 From list to powerset

We have seen that lists and subsets behave in a similar manner. We strengthen this connection by defining a *support* function $supp\colon \mathcal{L}(X) \to \mathcal{P}_{fin}(X)$ between them, via:

$$supp([x_1, \ldots, x_n]) \coloneqq \{x_1, \ldots, x_n\}.$$

Thus, the support of a list is the subset of elements occurring in the list. The support function removes order and multiplicities. The latter happens implicitly, via the set notation, above on the right-hand side. For instance,

$$supp([b, a, b, b, b]) = \{a, b\} = \{b, a\}.$$

Notice that there is no way to go in the other direction, namely $\mathcal{P}_{fin}(X) \to \mathcal{L}(X)$. Of course, one can for each subset *choose* an order of the elements in order to turn the subset into a list. However, this process is completely arbitrary and is not uniform (natural).

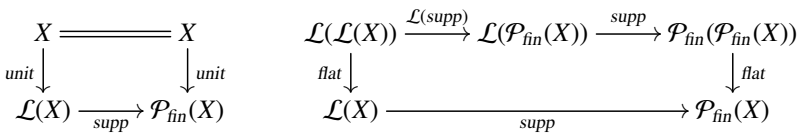
The support function interacts nicely with the structures that we have seen so far. This is expressed in the result below, where we use the same notation *unit* and *flat* for different functions, namely for $\mathcal{L}$ and for $\mathcal{P}$. The context, and especially the type of an argument, will make clear which one is meant.

Lemma 1.3.3. *Consider the support map* $supp\colon \mathcal{L}(X) \to \mathcal{P}_{fin}(X)$ *defined above.*

1 It is a map of monoids $(\mathcal{L}(X), [], +\!\!+) \to (\mathcal{P}(X), \emptyset, \cup)$.
2 It is natural, in the sense that for $f\colon X \to Y$ *one has:*

$$\begin{array}{ccc} \mathcal{L}(X) & \xrightarrow{supp} & \mathcal{P}_{fin}(X) \\ {\scriptstyle \mathcal{L}(f)}\downarrow & & \downarrow{\scriptstyle \mathcal{P}_{fin}(f)} \\ \mathcal{L}(Y) & \xrightarrow[supp]{} & \mathcal{P}_{fin}(Y) \end{array}$$

3 It commutes with the unit and flatten maps of list and powerset, as in:



Proof. The first item is easy and skipped. For item 2,

$$\begin{aligned}(\mathcal{P}_{fin}(f) \circ supp)([x_1, \ldots, x_n]) &= \mathcal{P}_{fin}(f)(supp([x_1, \ldots, x_n])) \\ &= \mathcal{P}_{fin}(f)(\{x_1, \ldots, x_n\}) \\ &= \{f(x_1), \ldots, f(x_n)\} \\ &= supp([f(x_1), \ldots, f(x_n)]) \\ &= supp(\mathcal{L}(f)([x_1, \ldots, x_n])) \\ &= (supp \circ \mathcal{L}(f))([x_1, \ldots, x_n]).\end{aligned}$$

In item 3, commutation of the first diagram is easy:

$$(supp \circ unit)(x) = supp([x]) = \{x\} = unit(x).$$

The second diagram requires a bit more work. Starting from a list of lists we get:

$$\begin{aligned}&(flat \circ supp \circ \mathcal{L}(supp))([[x_{11}, \ldots, x_{1n_1}], \ldots, [x_{k1}, \ldots, x_{kn_k}]]) \\ &= (\bigcup \circ supp)([supp([x_{11}, \ldots, x_{1n_1}]), \ldots, supp([x_{k1}, \ldots, x_{kn_k}])]) \\ &= (\bigcup \circ supp)([\{x_{11}, \ldots, x_{1n_1}\}, \ldots, \{x_{k1}, \ldots, x_{kn_k}\}]) \\ &= \bigcup(\{\{x_{11}, \ldots, x_{1n_1}\}, \ldots, \{x_{k1}, \ldots, x_{kn_k}\}\}) \\ &= \{x_{11}, \ldots, x_{1n_1}, \ldots, x_{k1}, \ldots, x_{kn_k}\} \\ &= supp([x_{11}, \ldots, x_{1n_1}, \ldots, x_{k1}, \ldots, x_{kn_k}]) \\ &= (supp \circ flat)([[x_{11}, \ldots, x_{1n_1}], \ldots, [x_{k1}, \ldots, x_{kn_k}]]).\end{aligned}$$

❑

1.3.2 Finite powersets and idempotent commutative monoids

We briefly review the relation between monoids and (finite) powersets. At an abstract level the situation is much like for lists, as described in Subsection 1.2.1. For instance, the commutative idempotent monoids $\mathcal{P}_{fin}(X)$ are free, like lists, in Proposition 1.2.3.

Proposition 1.3.4. *Let X be a set and $(M, 0, +)$ a commutative idempotent monoid, with a function $f\colon X \to M$ between them. Then there is a unique homomorphism of monoids $\overline{f}\colon (\mathcal{P}_{fin}(X), \emptyset, \cup) \to (M, 0, +)$ with $\overline{f} \circ unit = f$. We represent this situation in the diagram below.*

$$\begin{array}{ccc} X & \xrightarrow{\ unit\ } & \mathcal{P}_{fin}(X) \\ & \searrow^{f} & \downarrow{\scriptstyle \overline{f},\ homomorphism} \\ & & M \end{array} \tag{1.9}$$

Proof. Given the requirements, the only way to define $\overline{f}$ is as:

$$\overline{f}\big(\{x_1, \ldots, x_n\}\big) \coloneqq f(x_1) + \cdots + f(x_n), \qquad \text{with special case} \qquad \overline{f}(\emptyset) = 0.$$

The order in the above sum $f(x_1) + \cdots + f(x_n)$ does not matter since M is commutative. The function $\overline{f}$ sends unions to sums since + is idempotent. ❑

Commutative idempotent monoids can be described as algebras, in analogy with Proposition 1.2.6.

Proposition 1.3.5. *Let X be an arbitrary set.*

1 *To specify a commutative idempotent monoid structure $(u, +)$ on X is the same as giving a $\mathcal{P}_{fin}$-*algebra *$\alpha\colon \mathcal{P}_{fin}(X) \to X$, namely so that the diagrams*

$$\begin{array}{ccc} X & \xrightarrow{unit} & \mathcal{P}_{fin}(X) \\ & \searrow^{id} & \downarrow{\scriptstyle \alpha} \\ & & X \end{array} \qquad\qquad \begin{array}{ccc} \mathcal{P}_{fin}(\mathcal{P}_{fin}(X)) & \xrightarrow{\mathcal{P}_{fin}(\alpha)} & \mathcal{P}_{fin}(X) \\ {\scriptstyle flat}\downarrow & & \downarrow{\scriptstyle \alpha} \\ \mathcal{P}_{fin}(X) & \xrightarrow{\alpha} & X \end{array} \tag{1.10}$$

commute.

2 *Let $(M_1, u_1, +_1)$ and $(M_2, u_2, +_2)$ be two commutative idempotent monoids, with corresponding $\mathcal{P}_{fin}$-algebras $\alpha_1\colon \mathcal{P}_{fin}(M_1) \to M_1$ and $\alpha_2\colon \mathcal{P}_{fin}(M_2) \to M_2$. A function $f\colon M_1 \to M_2$ is a map of monoids if and only if the rectangle*

$$\begin{array}{ccc} \mathcal{P}_{fin}(M_1) & \xrightarrow{\mathcal{P}_{fin}(f)} & \mathcal{P}_{fin}(M_2) \\ {\scriptstyle \alpha_1}\downarrow & & \downarrow{\scriptstyle \alpha_2} \\ M_1 & \xrightarrow{f} & M_2. \end{array} \tag{1.11}$$

commutes.

Proof. This works very much like in the proof of Proposition 1.2.6. If $(X, u, +)$ is a monoid, we define $\alpha\colon \mathcal{P}_{fin}(X) \to X$ by freeness as $\alpha(\{x_1, \ldots, x_n\}) \coloneqq x_1 + \cdots + x_n$. In the other direction, given $\alpha\colon \mathcal{P}_{fin}(X) \to X$ we define a sum as $x + y \coloneqq \alpha(\{x, y\})$ with unit $u \coloneqq \alpha(\emptyset)$. Clearly, thi sum + on X is commutative and idempotent. ❑

This result concentrates on the finite powerset functor $\mathcal{P}_{fin}$. One can also consider algebras $\mathcal{P}(X) \to X$ for the (general) powerset functor $\mathcal{P}$. Such algebras turn the set X into a complete lattice, see [116, 9, 92] for details.

1.3.3 Extraction

So far we have concentrated on how similar lists and subsets are: the only structural difference that we have seen up to now is that subsets form an *idempotent*

and *commutative* monoid. But there are other important differences. Here we look at subsets of product sets, also known as relations.

The observation is that one can extract functions from a relation $R \subseteq X \times Y$, namely functions of the form $extr_1(R)\colon X \to \mathcal{P}(Y)$ and $extr_2(R)\colon Y \to \mathcal{P}(X)$, given by:

$$extr_1(R)(x) = \{y \in Y \mid (x, y) \in R\} \qquad extr_2(R)(y) = \{x \in X \mid (x, y) \in R\}.$$

In fact, one can easily reconstruct the relation R from $extr_1(R)$, and also from $extr_2(R)$, via:

$$R = \{(x, y) \mid y \in extr_1(R)(x)\} = \{(x, y) \mid x \in extr_2(R)(y)\}.$$

This all looks rather trivial, but such function extraction is less trivial for other data types, as we shall see later on for distributions, where it will be called disintegration.

Using the exponent notation from Subsection 1.1.2 we can summarise the situation as follows. There are two isomorphisms:

$$\mathcal{P}(Y)^X \cong \mathcal{P}(X \times Y) \cong \mathcal{P}(X)^Y. \tag{1.12}$$

Functions of the form $X \to \mathcal{P}(Y)$ will later be called 'channels' from X to Y, see Section 1.8. What we have just seen will then be described in terms of 'extraction of channels'.

1.3.4 Powerset combinatorics

We describe the basic of counting subsets, of finite sets. This will lead to the binomial and multinomial coefficients that show up in many places in probability theory.

Definition 1.3.6. For a finite set A we write $|A| \in \mathbb{N}$ for the number of elements in A. Then, for an arbitrary set X and a number $K \in \mathbb{N}$ we define:

$$\mathcal{P}[K](X) := \{U \in \mathcal{P}_{fin}(X) \mid |U| = K\}.$$

For a number $n \in \mathbb{N}$ we also write n for the n-element set $\{0, 1, \ldots, n-1\}$. Thus 0 is a also the empty set, and $1 = \{0\}$ is a singleton.

We recall that from a categorical perspective the sets 0 and 1 are special, because they are, respectively, initial and final in the category of sets and functions. We have $|A \times B| = |A| \cdot |B|$ and $|A + B| = |A| + |B|$, where $A + B$ is the disjoint union (coproduct) of sets A and B. It is well known that $|\mathcal{P}(A)| = 2^{|A|}$ for a finite set A. Below we are interested in the number of subsets of a fixed size.

The next lemma makes a basic result explicit, partly because it provides valuable insight in itself, but also because we shall see generalisations later on, for multisets instead of subsets. We use the familiar *binomial* and *multinomial* coefficients. We recall their definitions, for natural numbers $k \leq n$ and $m_1, \ldots, m_\ell$ with $\ell \geq 2$ and $\sum_i m_i = m$.

$$\binom{n}{k} := \frac{n!}{k! \cdot (n-k)!} \qquad \text{and} \qquad \binom{m}{m_1, \ldots, m_\ell} := \frac{m!}{m_1! \cdot \ldots \cdot m_\ell!}. \tag{1.13}$$

Recall that a *partition* of a set X is a disjoint cover: a finite collection of subsets $U_1, \ldots, U_k \subseteq X$ satisfying $U_i \cap U_j = \emptyset$ for $i \neq j$ and $\bigcup_i U_i = X$. We do not assume that the subsets U_i are non-empty.

Lemma 1.3.7. *Let X be a (finite) set with n elements.*

1 *The binomial coefficient counts the number of subsets of size $K \leq n$,*

$$\Big|\mathcal{P}[K](X)\Big| = \binom{n}{K}, \quad \textit{and especially:} \quad \Big|\mathcal{P}[K](n)\Big| = \binom{n}{K}.$$

2 *For $K_1, \ldots, K_\ell \in \mathbb{N}$ with $\ell \geq 2$ and $n = \sum_i K_i$, the multinomial coefficient counts the number of partitions, of appropriate sizes:*

$$\Big|\{\, \vec{U} \in \mathcal{P}[K_1](X) \times \cdots \times \mathcal{P}[K_\ell](X) \,\Big|\, U_1, \ldots, U_\ell \textit{ is a partition of } X \,\}\Big| = \binom{n}{K_1\ \ldots, K_\ell}.$$

Each subset $U \subseteq X$ forms a partition $(U, \neg U)$ of X, with its complement $\neg U = X \setminus U$. Correspondingly the binomial coefficient can be described as a multinomial coefficient:

$$\binom{n}{K} = \binom{n}{K, n-K}.$$

Proof. 1 We use induction on the number n of elements in X. If $n = 0$, also $K = 0$ and so $\mathcal{P}[K](X)$ contains only the empty set. At the same time, also $\binom{0}{0} = 1$. Next, let X have $n+1$ elements, say $X = Y \cup \{y\}$ where Y has n elements and $y \notin Y$. If $K = 0$ or $K = n+1$, there is only one subset in X of size K, and indeed $\binom{n+1}{0} = 1 = \binom{n+1}{n+1}$. We may thus assume $0 < K \leq n$. A subset of size K in X is thus either a subset of Y, or it contains y, and is then determined by a subset of size $K-1$. Hence:

$$\begin{aligned} \Big|\mathcal{P}[K](X)\Big| &= \Big|\mathcal{P}[K](Y)\Big| + \Big|\mathcal{P}[K-1](Y)\Big| \\ &\overset{\text{(IH)}}{=} \binom{n}{K} + \binom{n}{K-1} = \binom{n+1}{K}, \qquad \text{by Exercise 1.3.5.} \end{aligned}$$

2 We now use induction on $m \geq 2$. The case $m = 2$ has just been covered. Next:

$$\begin{aligned}
&\left|\{\, \vec{U} \in \mathcal{P}[K_1](X) \times \cdots \times \mathcal{P}[K_{m+1}](X) \,\middle|\, U_1, \ldots, U_{m+1} \text{ is a partition of } X \,\}\right| \\
&= \left|\{\, \vec{U} \in \mathcal{P}[K_1](X) \times \cdots \times \mathcal{P}[K_{m+1}](X) \,\middle|\, \right. \\
&\qquad\qquad \left. U_2, \ldots, U_{m+1} \text{ is a partition of } X \setminus U_1 \,\}\right| \\
&\overset{\text{(IH)}}{=} \binom{n}{K_1} \cdot \binom{n - K_1}{K_2, \ldots, K_{m+1}} \\
&= \binom{n}{K_1 \ \ldots, K_{m+1}}, \qquad \text{by Exercise 1.3.6.}
\end{aligned}$$

❑

The following result will be useful later.

Lemma 1.3.8. *Fix a number $m \in \mathbb{N}$. Then:*

$$\lim_{n\to\infty} \frac{\binom{n}{m}}{n^m} = \frac{1}{m!}.$$

Proof. Since:

$$\begin{aligned}
\lim_{n\to\infty} \frac{\binom{n}{m}}{n^m} &= \lim_{n\to\infty} \frac{n!}{m! \cdot (n-m)! \cdot n^m} \\
&= \frac{1}{m!} \cdot \lim_{n\to\infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{n-m+1}{n} \\
&= \frac{1}{m!} \cdot \left(\lim_{n\to\infty} \frac{n}{n}\right) \cdot \left(\lim_{n\to\infty} \frac{n-1}{n}\right) \cdot \ldots \cdot \left(\lim_{n\to\infty} \frac{n-m+1}{n}\right) \\
&= \frac{1}{m!}.
\end{aligned}$$

❑

Exercises

1.3.1 Continuing Exercise 1.2.1, compute:

- $\mathit{supp}(\ell_1)$
- $\mathit{supp}(\ell_2)$
- $\mathit{supp}(\ell_1 +\!\!+ \ell_2)$
- $\mathit{supp}(\ell_1) \cup \mathit{supp}(\ell_2)$
- $\mathit{supp}(\mathcal{L}(f)(\ell_1))$
- $\mathcal{P}_{\text{fin}}(f)(\mathit{supp}(\ell_1))$.

1.3.2 We have used finite unions $(\emptyset, \cup)$ as monoid structure on $\mathcal{P}(X)$ in Lemma 1.3.1 (1). Intersections $(X, \cap)$ give another monoid structure on $\mathcal{P}(X)$.

1 Show that the negation/complement function $\neg\colon \mathcal{P}(X) \to \mathcal{P}(X)$, given by:

$$\neg U = X \setminus U = \{x \in X \mid x \notin U\},$$

is a homomorphism of monoids between $(\mathcal{P}(X), \emptyset, \cup)$ and $(\mathcal{P}, X, \cap)$, in both directions. In fact, if forms an isomorphism of monoids, since $\neg\neg U = U$.

2 Prove that the intersections monoid structure is *not* preserved by maps $\mathcal{P}(f)\colon \mathcal{P}(X) \to \mathcal{P}(Y)$.
Hint: Look at preservation of the unit $X \in \mathcal{P}(X)$.

1.3.3 Check that in general,

$$\big|\mathcal{P}(f)(U)\big| \neq \big|U\big|.$$

Conclude that the powerset functor $\mathcal{P}$ does not restrict to a functor $\mathcal{P}[K]$, for $K \in \mathbb{N}$.

1.3.4 Check that an ordered partition of a set X of size m can be identified with a function $X \to m$. This assumes that the partition consists of non-empty subsets.

1.3.5 1 Prove what is called Pascal's rule: for $0 < m \leq n$,

$$\binom{n+1}{m} = \binom{n}{m} + \binom{n}{m-1}. \tag{1.14}$$

2 Use this equation to prove:

$$\sum_{0 \leq i \leq n} \binom{m+i}{m} = \binom{n+m+1}{m+1}.$$

3 Now use both previous equations to prove:

$$\sum_{0 \leq i \leq n} (n+1-i) \cdot \binom{m+i}{m} = \binom{n+m+2}{m+2}.$$

1.3.6 Let $K_1, \ldots, K_\ell \in \mathbb{N}$ be given, for $\ell > 2$, with $K = \sum_i K_i$. Show that the multinomial coefficient can be written as a product of binomial coefficients, via:

$$\binom{K}{K_1, \ldots, K_\ell} = \binom{K}{K_1} \cdot \binom{K-K_1}{K_2, \ldots, K_\ell}$$

1.3.7 Let $K_1, \ldots, K_\ell \in \mathbb{N}$ be given, for $\ell \geq 2$, with $K = \sum_i K_i$. Show that:

$$\Big|\mathcal{P}[K_1](K) \times \mathcal{P}[K_2](K-K_1) \times \cdots \times \mathcal{P}[K_{\ell-1}](K-K_1-\cdots-K_{\ell-2})\Big| = \binom{K}{K_1\ \ldots, K_\ell}.$$

1.4 Multisets

So far we have discussed two collection data types, namely lists and subsets of elements. In lists, elements occur in a particular order, and may occur multiple times (at different positions). Both properties are lost when moving from lists to sets. In this section we look at multisets, which are 'sets' in which elements may occur multiple times. Hence multisets are in between lists and subsets, since they do allow multiple occurrences, but the order of their elements is irrelevant.

The list and powerset examples are somewhat remote from probability theory. But multisets are much more directly relevant: first, because we use a similar notation for multisets and distributions; and second, because observed data can be organised nicely in terms of multisets. For instance, for statistical analysis, a document is often seen as a multiset of words, in which one keeps track of the words that occur in the document together with their frequency (multiplicity); in that case, the order of the words is ignored. Also, tables with observed data can be organised naturally as multisets, see Subsection 1.4.1 below. Learning from such tables will be described in Section 2.1 as a (natural) transformation from multisets to distributions.

Despite their importance, multisets do not have a prominent role in programming, like lists have. Eigenvalues of a matrix form a clear example where the 'multi' aspect is ignored in mathematics: eigenvalues may occur multiple times, so the proper thing to say is that a matrix has a *multiset* of eigenvalues. One reason for not using multisets may be that there is no established notation. We shall use a 'ket' notation $| - \rangle$ that is borrowed from quantum theory, but interchangeably also a functional notation. Since multisets are less familiar, we take time to introduce the basic definitions and properties, in Sections 1.4 – 1.7.

We start with an introduction about notation, terminology, and conventions for multisets. Consider a set $C = \{R, G, B\}$ for the three colours Red, Green, Blue. An example of a multiset over C is:

$$2|R\rangle + 5|G\rangle + 0|B\rangle.$$

In this multiset the element R occurs 2 times, G occurs 5 times, and B occurs 0 times. The latter means that B does not occur, that is, B is not an element of the multiset. From a multiset perspective, we have $2 + 5 + 0 = 7$ elements — and not just 2. A multiset like this may describe an urn containing 2 red balls, 5 green ones, and no blue balls. Such multisets are quite common. For instance, the chemical formula $C_2H_3O_2$ for vinegar may be read as a multiset

$2|C\rangle + 3|H\rangle + 2|O\rangle$, containing 2 carbon (C) atoms, 3 hydrogen (H) atoms and 2 oxygen (O) atoms.

In a situation where we have multiple data items, say arising from successive experiments, a basic question to ask is: does the order of the experiments matter? If so, we need to order the data elements as a list. If the order does not matter we should use a multiset. More concretely, if six successive experiments yield data items d, e, d, f, e, d and their order is relevant we should model the data as the list $[d, e, d, f, e, d]$. When the order is irrelevant, we can capture the data as the multiset $3|d\rangle + 2|e\rangle + 1|f\rangle$.

The funny brackets $|-\rangle$ are called *ket* notation; this is frequently used in quantum theory. Here it is meaningless notation, used to separate the natural numbers, called *multiplicities*, and the elements in the multiset.

Let X be an arbitrary set. Following the above ket-notation, a (finite) multiset over X is an expression of the form:

$$n_1|x_1\rangle + \cdots + n_k|x_k\rangle \qquad \text{where} \qquad n_i \in \mathbb{N} \text{ and } x_i \in X.$$

This expression is a formal sum, not an actual sum (for instance in $\mathbb{R}$). We may write it as $\sum_i n_i|x_i\rangle$. We use the convention:

- $0|x\rangle$ may be omitted; but it may also be written explicitly in order to emphasise that the element x does not occur in a multiset;
- a sum $n|x\rangle + m|x\rangle$ is the same as $(n+m)|x\rangle$;
- the order and brackets (if any) in a sum do not matter.

Thus, for instance, there is an equality of multisets:

$$2|a\rangle + (5|b\rangle + 0|c\rangle) + 4|b\rangle \;=\; 9|b\rangle + 2|a\rangle.$$

There is an alternative description of multisets. A multiset can be defined as a function $\varphi\colon X \to \mathbb{N}$ that has finite support. The support $\mathit{supp}(\varphi) \subseteq X$ is the set $\mathit{supp}(\varphi) = \{x \in X \mid \varphi(x) \neq 0\}$. For each element $x \in X$ the number $\varphi(x) \in \mathbb{N}$ indicates how often x occurs in the multiset φ. Such a function φ can also be written as a formal sum $\sum_x \varphi(x)|x\rangle$, where x ranges over $\mathit{supp}(\varphi)$.

For instance, the multiset $9|b\rangle + 2|a\rangle$ over $A = \{a, b, c\}$ corresponds to the function $\varphi\colon A \to \mathbb{N}$ given by $\varphi(a) = 2$, $\varphi(b) = 9$, $\varphi(c) = 0$. Its support is thus $\{a, b\} \subseteq A$, with two elements. The number $\|\varphi\|$ of elements in φ is 11.

We shall freely switch back and forth between the ket-description and the function-description of multisets, and use whichever form is most convenient for the goal at hand.

Having said this, we stretch the idea of a multiset and do not only allow natural numbers $n \in \mathbb{N}$ as multiplicities, but also allow non-negative numbers $r \in \mathbb{R}_{\geq 0}$. Thus we can have a multiset of the form $\frac{3}{2}|a\rangle + \pi|b\rangle$ where $\pi \in \mathbb{R}_{\geq 0}$

is the famous constant of Archimedes: the ratio of a circle's circumference to its diameter. This added generality will be useful at times, although many examples of multisets will simply have natural numbers as multiplicities. We call such multisets *natural*.

Definition 1.4.1. For a set X we shall write $\mathcal{M}(X)$ for the set of all *multisets* over X. Thus, using the function approach:

$$\mathcal{M}(X) \coloneqq \{\varphi \colon X \to \mathbb{R}_{\geq 0} \mid \mathit{supp}(\varphi) \text{ is finite}\}.$$

The elements of $\mathcal{M}(X)$ may be called *mass functions*, as in [147].

We shall write $\mathcal{N}(X) \subseteq \mathcal{M}(X)$ for the subset of *natural* multisets, with natural numbers as multiplicities — also called *bags* or *urns*. Thus, $\mathcal{N}(X)$ contains functions $\varphi \in \mathcal{M}(X)$ with $\varphi(x) \in \mathbb{N}$, for all $x \in X$.

We shall write $\mathcal{M}_*(X)$ for the set of *non-empty* multisets. Thus:

$$\begin{aligned} \mathcal{M}_*(X) &\coloneqq \{\varphi \in \mathcal{M}(X) \mid \mathit{supp}(\varphi) \text{ is non-empty}\} \\ &= \{\varphi \colon X \to \mathbb{R}_{\geq 0} \mid \mathit{supp}(\varphi) \text{ is finite and non-empty}\}. \end{aligned}$$

Similarly, $\mathcal{N}_*(X) \subseteq \mathcal{M}_*(X)$ contains the non-empty natural multisets.

For a number K we shall write $\mathcal{M}[K](X) \subseteq \mathcal{M}(X)$ and $\mathcal{N}[K](X) \subseteq \mathcal{N}(X)$ for the subsets of mulisets with K elements. Thus:

$$\mathcal{M}[K](X) \coloneqq \{\varphi \in \mathcal{M}(X) \mid \|\varphi\| = K\} \qquad \text{where} \qquad \|\varphi\| \coloneqq \textstyle\sum_x \varphi(x). \tag{1.15}$$

This expression $\|\varphi\|$ gives the *size* of the multiset, that is, its total number of elements.

All of $\mathcal{M}, \mathcal{N}, \mathcal{M}_*, \mathcal{N}_*, \mathcal{M}[K], \mathcal{N}[K]$ are functorial, in the same way. Hence we concentrate on $\mathcal{M}$. For a function $f \colon X \to Y$ we can define $\mathcal{M}(f) \colon \mathcal{M}(X) \to \mathcal{M}(Y)$. When we see a multiset $\varphi \in \mathcal{M}(X)$ as an urn containing coloured balls, with colours from X, then $\mathcal{M}(f)(\varphi) \in \mathcal{M}(Y)$ is the urn with 'repainted' balls, where the new colours are taken from the set Y. The function $f \colon X \to Y$ defines the transformation of colours. It tells that a ball of colour $x \in X$ in φ should be repainted with colour $f(x) \in Y$.

The urn $\mathcal{M}(f)(\varphi)$ with repainted balls can be defined in two equivalent ways:

$$\mathcal{M}(f)\Big(\textstyle\sum_i r_i|x_i\rangle\Big) \coloneqq \sum_i r_i|f(x_i)\rangle \qquad \text{or as} \qquad \mathcal{M}(f)(\varphi)(y) \coloneqq \sum_{x \in f^{-1}(y)} \varphi(x).$$

It may take a bit of effort to see that these two descriptions are the same, see Exercise 1.4.1 below. Notice that in the sum $\sum_i r_i|f(x_i)\rangle$ it may happen that $f(x_{i_1}) = f(x_{i_2})$ for $x_{i_1} \neq x_{i_2}$, so that r_{i_1} and r_{i_2} are added together. Thus, the support of $\mathcal{M}(f)(\sum_i r_i|x_i\rangle)$ may have fewer elements than the support of $\sum_i r_i|x_i\rangle$, but the sum of all multiplicities is the same in $\mathcal{M}(f)(\sum_i r_i|x_i\rangle)$ and $\sum_i r_i|x_i\rangle$, see Exercise 1.5.2 below.

Applying $\mathcal{M}$ to a projection function $\pi_i \colon X_1 \times \cdots \times X_n \to X_i$ yields a function $\mathcal{M}(\pi_i)$, from the set $\mathcal{M}(X_1 \times \cdots \times X_n)$ of multisets over a product to the set $\mathcal{M}(X_i)$ of multisets over a component. This $\mathcal{M}(\pi_i)$ is called a *marginal function* or simply a *marginal*. It computes what is 'on the side', in the marginal of a table, as will be illustrated in Subsection 1.4.1 below.

Multisets, like lists and subset form a monoid. In terms of urns with coloured balls, taking the sum of two multisets corresponds to pouring the balls from two urns into a new urn.

Lemma 1.4.2. *1 The set $\mathcal{M}(X)$ of multisets over X is a commutative monoid. In functional form, addition and zero (identity) element $\mathbf{0} \in \mathcal{M}(X)$ are defined as:*

$$(\varphi + \psi)(x) \coloneqq \varphi(x) + \psi(x) \qquad\qquad \mathbf{0}(x) \coloneqq 0.$$

These sums restrict to $\mathcal{N}(X)$.

2 The set $\mathcal{M}(X)$ is also a cone*: it is closed under 'scalar' multiplication with non-negative numbers $r \in \mathbb{R}_{\geq 0}$, via:*

$$(r \cdot \varphi)(x) \coloneqq r \cdot \varphi(x).$$

This scalar multiplication $r \cdot (-) \colon \mathcal{M}(X) \to \mathcal{M}(X)$ preserves the sums $(\mathbf{0}, +)$ from the previous item, and is thus a map of monoids.

3 For each $f \colon X \to Y$, the function $\mathcal{M}(f) \colon \mathcal{M}(X) \to \mathcal{M}(Y)$ is a map of monoids and also of cones. The latter means: $\mathcal{M}(f)(r \cdot \varphi) = r \cdot \mathcal{M}(f)(\varphi)$. ❑

The fact that $\mathcal{M}(f)$ preserves sums can be understood informally as follows. If we have two urns, we can first combine their contents and then repaint everything. Alternatively, we can first repaint the balls in the two urns separately, and then throw them together. The result is the same, in both cases.

The element $\mathbf{0} \in \mathcal{M}(X)$ used in item (1) is the empty multiset, that is, the urn containing no balls. Similarly, the sum of multisets + is implicit in the ket-notation. The set $\mathcal{N}(X)$ of *natural* multisets is not closed in general under scalar multiplication with $r \in \mathbb{R}_{\geq 0}$. It is closed under scalar multiplication with $n \in \mathbb{N}$, but such multiplications add nothing new since they can also be described via repeated addition.

1.4.1 Tables of data as multisets

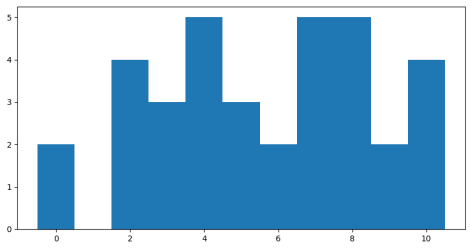
Let's assume that a group of 36 children in the age range $0 - 10$ is participating in some study, where the number of children of each age is given by the following table.

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	0	4	3	5	3	2	5	5	2	4

We can represent this table as a natural multiset over the set of ages $\{0, 1, \ldots, 10\}$.

$$2|0\rangle + 4|2\rangle + 3|3\rangle + 5|4\rangle + 3|5\rangle + 2|6\rangle + 5|7\rangle + 5|8\rangle + 2|9\rangle + 4|10\rangle.$$

Notice that there is no summand for age 1 because of our convention to ommit expressions like $0|1\rangle$ with multiplicity 0. We can visually represent the above age data/multiset in the form of a histogram:



(1.16)

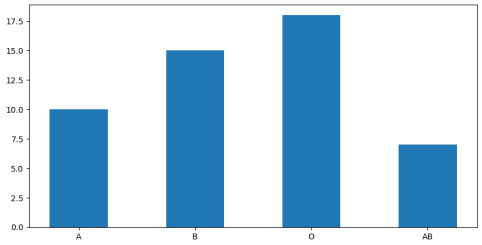
Here is another example, not with numerical data, in the form of ages, but with nominal data, in the form of blood types. Testing the blood type of 50 individuals gives the following table.

A	**B**	**O**	**AB**
10	15	18	7

This corresponds to a (natural) multiset over the set $\{A, B, O, AB\}$ of blood types, namely to:

$$10|A\rangle + 15|B\rangle + 18|O\rangle + 7|AB\rangle.$$

It gives rise to the following bar graph, in which there is no particular ordering of elements. For convenience, we follow the order of the above table.



(1.17)

Next, consider the two-dimensional table (1.18) below where we have combined numeric information about blood pressure (either high H, or low L) and certain medicines (either type 1, type 2, or no medicine, indicated as 0). There is data about 100 study participants:

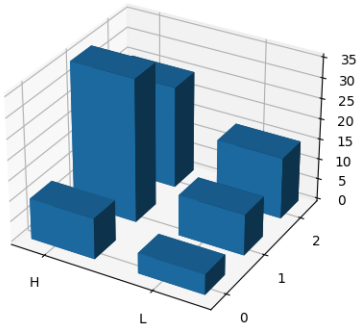
	no medicine	**medicine 1**	**medicine 2**	**totals**
high	10	35	25	70
low	5	10	15	30
totals	15	45	40	100

(1.18)

We claim that we can capture this table as a (natural) multiset. To do so, we first form sets $B = \{H, T\}$ for blood pressure values, and $M = \{0, 1, 2\}$ for types of medicine. The above table can then be described as a natural multiset τ over the product set/space $B \times M$, that is, as an element $\tau \in \mathcal{N}(B \times M)$, namely:

$$\tau = 10|H, 0\rangle + 35|H, 1\rangle + 25|H, 2\rangle + 5|L, 0\rangle + 10|L, 1\rangle + 15|L, 2\rangle.$$

Such a multiset can be plotted in three dimensions as:



We see that Table (1.18) contains 'totals' in its vertical and horizontal margins. They can be obtained from τ as marginals, using the functoriality of $\mathcal{N}$. This works as follows. Applying the natural multiset functor $\mathcal{N}$ to the two projections $\pi_1 \colon B \times M \to B$ and $\pi_2 \colon B \times M \to M$ yields marginal distributions on

B and M, namely:

$$\begin{aligned}
\mathcal{N}(\pi_1)(\tau) &= 10|\pi_1(H,0)\rangle + 35|\pi_1(H,1)\rangle + 25|\pi_1(H,2)\rangle \\
&\qquad + 5|\pi_1(L,0)\rangle + 10|\pi_1(L,1)\rangle + 15|\pi_1(L,2)\rangle \\
&= 10|H\rangle + 35|H\rangle + 25|H\rangle + 5|L\rangle + 10|L\rangle + 15|L\rangle \\
&= 70|H\rangle + 30|L\rangle. \\
\mathcal{N}(\pi_2)(\tau) &= (10+5)|0\rangle + (35+10)|1\rangle + (25+15)|2\rangle \\
&= 15|0\rangle + 45|1\rangle + 40|2\rangle.
\end{aligned}$$

The expression 'marginal' is used to describe such totals in the margin of a multidimensional table. In Section 2.1 we describe how to obtain probabilities from tables in a systematic manner.

1.4.2 Unit and flatten for multisets

As may be expected by now, there are also unit and flatten maps for multisets. The unit function $\mathit{unit}\colon X \to \mathcal{M}(X)$ is simply $\mathit{unit}(x) \coloneqq 1|x\rangle$. Flattening involves turning a multiset of multisets into a multiset. Concretely, this is done as:

$$\mathit{flat}\Big(\tfrac{1}{3}\Big|\, 2|a\rangle + 2|c\rangle \,\Big\rangle + 5\Big|\, 1|b\rangle + \tfrac{1}{6}|c\rangle \,\Big\rangle\Big) = \tfrac{2}{3}|a\rangle + 5|b\rangle + \tfrac{3}{2}|c\rangle.$$

More generally, flattening is the function $\mathit{flat}\colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$ with:

$$\mathit{flat}\Big(\textstyle\sum_i r_i|\varphi_i\rangle\Big) \coloneqq \displaystyle\sum_x \big(\textstyle\sum_i r_i \cdot \varphi_i(x)\big)\,\big|x\rangle.$$

Notice that the big outer sum $\sum_x$ is a formal one, whereas the inner sum $\sum_i$ is an actual one, in $\mathbb{R}_{\geq 0}$, see the earlier example.

The following result, about unit and flatten, does not come as a surprise anymore. We formulate it for general multisets $\mathcal{M}$, but it restricts to natural multisets $\mathcal{N}$.

Lemma 1.4.3. *1 For each function $f\colon X \to Y$ the two rectangles*

$$\begin{array}{ccc} X & \xrightarrow{\mathit{unit}} & \mathcal{M}(X) \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle \mathcal{M}(f)} \\ Y & \xrightarrow[\mathit{unit}]{} & \mathcal{M}(Y) \end{array} \qquad\qquad \begin{array}{ccc} \mathcal{M}(\mathcal{M}(X)) & \xrightarrow{\mathit{flat}} & \mathcal{M}(X) \\ {\scriptstyle \mathcal{M}(\mathcal{M}(f))}\downarrow & & \downarrow{\scriptstyle \mathcal{M}(f)} \\ \mathcal{M}(\mathcal{M}(Y)) & \xrightarrow[\mathit{flat}]{} & \mathcal{M}(Y) \end{array}$$

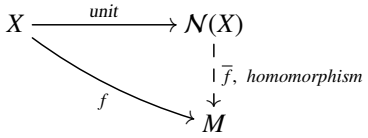
commute.

2 The next two diagrams also commute.

$$\begin{array}{ccccc} \mathcal{M}(X) & \xrightarrow{\mathit{unit}} & \mathcal{M}(\mathcal{M}(X)) & \xleftarrow{\mathcal{M}(\mathit{unit})} & \mathcal{M}(X) \\ & \searrow & \downarrow{\scriptstyle \mathit{flat}} & \swarrow & \\ & & \mathcal{M}(X) & & \end{array} \qquad \begin{array}{ccc} \mathcal{M}(\mathcal{M}(\mathcal{M}(X))) & \xrightarrow{\mathit{flat}} & \mathcal{M}(\mathcal{M}(X)) \\ {\scriptstyle \mathcal{M}(\mathit{flat})}\downarrow & & \downarrow{\scriptstyle \mathit{flat}} \\ \mathcal{M}(\mathcal{M}(X)) & \xrightarrow[\mathit{flat}]{} & \mathcal{M}(X) \end{array} \qquad \square$$

The next result shows that *natural* multisets are free commutative monoids. Arbitrary multisets are also free, but for other algebraic structures, see Exercise 1.4.9.

Proposition 1.4.4. *Let X be a set and $(M, 0, +)$ a commutative monoid. Each function $f\colon X \to M$ has a unique extension to a homomorphism of monoids $\overline{f}\colon (\mathcal{N}(X), \mathbf{0}, +) \to (M, 0, +)$ with $\overline{f} \circ \mathsf{unit} = f$. The diagram below captures the situation, where the dashed arrow is used for uniqueness.*



(1.19)

Proof. One defines:

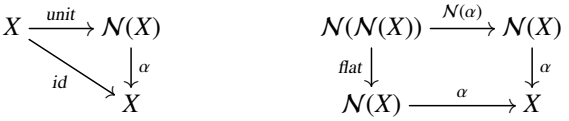
$$\overline{f}\Big(n_1|x_1\rangle + \cdots + n_k|x_k\rangle\Big) \coloneqq n_1 \cdot f(x_1) + \cdots + n_k \cdot f(x_k).$$

where we write $n \cdot a$ for the n-fold sum $a + \cdots + a$ in a monoid. ❑

The unit and flatten operations for (natural) multisets can be used to capture commutative monoids more precisely, in analogy with Propositions 1.2.6 and 1.3.5

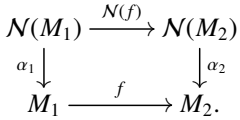
Proposition 1.4.5. *Let X be an arbitrary set.*

1 *A commutative monoid structure $(u, +)$ on X corresponds to an $\mathcal{N}$-*algebra *$\alpha\colon \mathcal{N}(X) \to X$ making the two diagrams below commute.*



(1.20)

2 *Let $(M_1, u_1, +_1)$ and $(M_2, u_2, +_2)$ be two commutative monoids, with corresponding $\mathcal{N}$-algebras $\alpha_1\colon \mathcal{N}(M_1) \to M_1$ and $\alpha_2\colon \mathcal{N}(M_2) \to M_2$. A function $f\colon M_1 \to M_2$ is a map of monoids if and only if the rectangle*



(1.21)

commutes.

Proof. Analogously to the proof Proposition 1.2.6: if $(X, u, +)$ is a commutative monoid, we define $\alpha\colon \mathcal{N}(X) \to X$ by turning formal sums into actual

sums: $\alpha(\sum_i n_i|x_i\rangle) \coloneqq \sum_i n_i \cdot x_i$, see (the proof of) Proposition 1.4.4. In the other direction, given $\alpha\colon \mathcal{N}(X) \to X$ we define a sum as $x+y \coloneqq \alpha(1|x\rangle + 1|y\rangle)$ with unit $u \coloneqq \alpha(\mathbf{0})$. Obviously, $+$ is commutative. ❑

1.4.3 Extraction

At the end of the previous section we have seen how to extract a function (channel) from a binary subset, that is, from a relation. It turns out that one can do the same for a binary multiset, that is, for a table. More specifically, in terms of exponents, there are isomorphisms:

$$\mathcal{M}(Y)^X \cong \mathcal{M}(X \times Y) \cong \mathcal{M}(X)^Y. \tag{1.22}$$

This is analogous to (1.12) for powerset.

How does this work in detail? Suppose we have an arbitary multiset/table $\sigma \in \mathcal{M}(X \times Y)$. From σ one can extract a function $extr_1(\sigma)\colon X \to \mathcal{M}(Y)$, and also $extr_2(\sigma)\colon Y \to \mathcal{M}(X)$, via a cap:

$$extr_1(\sigma)(x) = \sum_{y \in Y} \sigma(x,y)|y\rangle \qquad\qquad extr_2(\sigma)(y) = \sum_{x \in X} \sigma(x,y)|x\rangle.$$

Notice that we are — conveniently — mixing ket and function notation for multisets. Conversely, σ can be reconstructed from $extr_1(\sigma)$, and also from $extr_2(\sigma)$, via $\sigma(x,y) = extr_1(\sigma)(x)(y) = extr_2(\sigma)(y)(x)$.

Functions of the form $X \to \mathcal{M}(Y)$ will also be used as channels from X to Y, see Section 1.8. That's why we often speak about 'channel extraction'.

As illustration, we apply extraction to the medicine - blood pressure Table 1.18, described as the multiset $\tau \in \mathcal{M}(B \times M)$. It gives rise to two channels $extr_1(\tau)\colon B \to \mathcal{M}(M)$ and $extr_2(\tau)\colon M \to \mathcal{M}(B)$. Explicitly:

$$\begin{aligned} extr_1(\tau)(H) &= \sum_{x \in M} \tau(H,x)|x\rangle = 10|0\rangle + 35|1\rangle + 25|2\rangle \\ extr_1(\tau)(L) &= \sum_{x \in M} \tau(L,x)|x\rangle = 5|0\rangle + 10|1\rangle + 15|2\rangle. \end{aligned}$$

We see that this extracted function captures the two rows of Table 1.18. Similarly we get the columns via the second extracted function:

$$\begin{aligned} extr_2(\tau)(0) &= 10|L\rangle + 5|H\rangle \\ extr_2(\tau)(1) &= 35|L\rangle + 10|H\rangle \\ extr_2(\tau)(2) &= 25|L\rangle + 15|H\rangle. \end{aligned}$$

Exercises

1.4.1 In the setting of Exercise 1.2.1, consider the multisets $\varphi = 3|a\rangle + 2|b\rangle + 8|c\rangle$ and $\psi = 3|b\rangle + 1|c\rangle$. Compute:

- $\varphi + \psi$
- $\psi + \varphi$
- $\mathcal{M}(f)(\varphi)$, both in ket-formulation and in function-formulation
- idem for $\mathcal{M}(f)(\psi)$
- $\mathcal{M}(f)(\varphi + \psi)$
- $\mathcal{M}(f)(\varphi) + \mathcal{M}(f)(\psi)$.

1.4.2 Consider, still in the context of Exercise 1.2.1, the 'joint' multiset $\varphi \in \mathcal{M}(X \times Y)$ given by $\varphi = 2|a, u\rangle + 3|a, v\rangle + 5|c, v\rangle$. Determine the marginals $\mathcal{M}(\pi_1)(\varphi) \in \mathcal{M}(X)$ and $\mathcal{M}(\pi_2)(\varphi) \in \mathcal{M}(Y)$.

1.4.3 Consider the chemical equation for burning methane:

$$CH_4 \;+\; 2\,O_2 \;\longrightarrow\; CO_2 \;+\; 2\,H_2O.$$

Check that there is an underlying equation of multisets:

$$\begin{aligned} &\mathit{flat}\Big(1\Big|\,1|C\rangle + 4|H\rangle\Big\rangle + 2\Big|2|O\rangle\Big\rangle\Big) \\ &= \mathit{flat}\Big(1\Big|\,1|C\rangle + 2|O\rangle\Big\rangle + 2\Big|2|H\rangle + 1|O\rangle\Big\rangle\Big). \end{aligned}$$

It expresses the law of conservation of mass.

1.4.4 Check that $\mathcal{M}(id) = id$ and $\mathcal{M}(g \circ f) = \mathcal{M}(g) \circ \mathcal{M}(f)$.

1.4.5 Show that for each natural number K the mappings $X \mapsto \mathcal{M}[K](X)$ and $X \mapsto \mathcal{N}[K](X)$ are functorial. Notice the difference with $\mathcal{P}[K]$, see Exercise 1.3.3.

1.4.6 Prove Lemma 1.4.3.

1.4.7 Consider both Lemma 1.4.3 and Proposition 1.4.5.

1 Notice that an abstract way of seeing that $\mathcal{N}(X)$ is a commutative monoid is via the properties of the flatten map $\mathit{flat}\colon \mathcal{N}(\mathcal{N}(X)) \to \mathcal{N}(X)$.

2 Notice also that this flatten map is a homomorphism of monoids.

1.4.8 Verify that the support map $\mathit{supp}\colon \mathcal{M}(X) \to \mathcal{P}_{\mathit{fin}}(X)$ commutes with extraction functions, in the sense that the following diagram commutes.

$$\begin{array}{ccc} \mathcal{M}(X \times Y) & \xrightarrow{\mathit{supp}} & \mathcal{P}_{\mathit{fin}}(X \times Y) \\ {\scriptstyle \mathit{extr}_1}\downarrow & & \downarrow{\scriptstyle \mathit{extr}_1} \\ \mathcal{M}(Y)^X & \xrightarrow[\mathit{supp}^X]{} & \mathcal{P}_{\mathit{fin}}(Y)^X \end{array}$$

Equationally, this amounts to showing that for $\tau \in \mathcal{M}(X \times Y)$ and $x \in X$ one has:

$$extr_1(supp(\tau))(x) = supp(extr_1(\tau)(x)).$$

Here we use that $supp^X(f) \coloneqq supp \circ f$, so that $supp^X(f)(x) = supp(f(x))$.

1.4.9 In Proposition 1.4.4 we have seen that natural multisets $\mathcal{N}(X)$ form free commutative monoids. What about general multisets $\mathcal{M}(X)$? They form free cones. Briefly, a cone is a commutative monoid M with scalar multiplication $r \cdot (-) \colon M \to M$, for each $r \in \mathbb{R}_{\geq 0}$, that is a homomorphism of monoids. It is like a vector space, not over all reals, but only over the non-negative reals. Homomorphisms of cones preserve such scalar multiplications.

Let X be a set and $(M, 0, +)$ a cone, with a function $f \colon X \to M$. Prove that there is a unique homomorphism of cones $\overline{f} \colon \mathcal{M}(X) \to M$ with $\overline{f} \circ unit = f$.

1.5 Multisets in summations

The current and next section will dive deeper into the use of natural multisets, in combinatorics as preparation for later use. This section will focus on the use of multisets in summations, such as the multinomial theorem — which is probably best known in binary form, as the binomial theorem for expanding sums $(a + b)^n$. We shall describe various extensions, both for finite and infinite sums. All material in this section is standard, but its presentation in terms of multisets is not.

Recall that a multiset $\varphi = \sum_i r_i |x_i\rangle$ is called *natural* when all the multiplicities r_i are natural numbers. Such a multiset is also called a *bag* or an *urn*. Urns are often used as illustrations in simple probabilistic arguments, starting with: what is the probability of drawing two red balls — with or without replacement — from an urn with initially three red balls and two blue ones. In this book we shall describe such an urn as the multiset $3|R\rangle + 2|B\rangle$. In such arguments one often encounters elementary combinatorial questions about (combinations of) numbers of balls in urns. This section provides some standard results, building on Subsection 1.3.4, where subsets are counted — instead of multisets.

In this section we focus on counting *with* multisets, in particular in (infinite) sums of powers. The next section focuses on counting multisets themselves, where we ask, for instance, how many multisets of size K are there on a set with n elements?

There are several ways to associate a natural number with a multiset φ. For instance, we can look at the size of its support $|supp(\varphi)| \in \mathbb{N}$, or at its size, as total number of elements $\|\varphi\| = \sum_x \varphi(x) \in \mathbb{R}_{>0}$. This size is a natural number when φ is a natural multiset. Below we will introduce several more such numbers for natural multisets φ, namely $\varphi!$ and $\binom{\varphi}{}$, and later on also a binomial coefficient $\binom{\psi}{\varphi}$.

Definition 1.5.1. 1 For two multisets $\varphi, \psi \in \mathcal{N}(X)$ we write:

$$\varphi \leq \psi \iff \forall x \in X.\, \varphi(x) \leq \psi(x).$$

When $\varphi \leq \psi$ we define subtraction $\psi - \varphi$ of multisets as the obvious multiset, defined pointwise as: $(\psi - \varphi)(x) = \psi(x) - \varphi(x)$.

2 For an additional number $K \in \mathbb{N}$ we use:

$$\varphi \leq_K \psi \iff \varphi \in \mathcal{N}[K](X) \text{ and } \varphi \leq \psi.$$

3 For a collection of numbers $r = (r_x)_{x \in X}$ we write:

$$r^\varphi := \prod_{x \in X} r_x^{\varphi(x)} = \vec{r}^{\,\varphi}.$$

The latter vector notation is appropriate in a situation with a particular order.

4 The factorial $\varphi!$ of a natural multiset $\varphi \in \mathcal{N}(X)$ is the product of the factorial of its multiplicities:

$$\varphi! := \prod_{x \in supp(\varphi)} \varphi(x)! \tag{1.23}$$

5 The *muliset coefficient* $\binom{\varphi}{}$ is defined as:

$$\binom{\varphi}{} := \frac{\|\varphi\|!}{\varphi!} = \frac{\|\varphi\|!}{\prod_x \varphi(x)!} = \binom{\|\varphi\|}{\varphi(x_1), \ldots, \varphi(x_n)}.$$

The latter formulation, using the multinomial coefficient (1.13), assumes that the support of φ is ordered as $[x_1, \ldots, x_n]$.

For instance,

$$(3|R\rangle + 2|B\rangle)! = 3! \cdot 2! = 12 \quad \text{and} \quad \binom{3|R\rangle + 2|B\rangle}{} = \frac{5!}{12} = 10.$$

The multinomial coefficient $\binom{\varphi}{}$ in item (5) counts the number of ways of putting $\|\varphi\|$ items in $supp(\varphi) = \{x_1, \ldots, x_n\}$ urns, with the restriction that $\varphi(x_i)$ items go into urn x_i. Alternatively, $\binom{\varphi}{}$ is the numbers of partitions (U_i) of a set X with $\|\varphi\|$ elements, where $|U_i| = \varphi(x_i)$.

The traditional notation $\binom{N}{m_1, \ldots, m_k}$ for multinomial coefficients in (1.13) is

suboptimal for two reasons: first, the number N is superflous, since it is determined by the m_i as $N = \sum_i m_i$; second, the order of the m_i is irrelevant. These disadvantages are resolved by the multiset variant $\mathbf{(}\varphi\mathbf{)}$. It has our preference.

We recall the *recurrence relations*:

$$\binom{K-1}{k_1-1,\ldots,k_n} + \cdots + \binom{K-1}{k_1,\ldots,k_n-1} = \binom{K}{k_1,\ldots,k_n} \tag{1.24}$$

for multinomial coefficients. A snappy re-formulation, for a natural multiset φ, is:

$$\sum_{x \in supp(\varphi)} \mathbf{(}\varphi - 1|x\rangle\mathbf{)} = \mathbf{(}\varphi\mathbf{)}. \tag{1.25}$$

Multinomial coefficients (1.13) are useful, for instance in the *Multinomial Theorem* (see *e.g.* [145]):

$$(r_1 + \cdots + r_n)^K = \sum_{k_i,\, \sum_i k_i = K} \binom{K}{k_1,\ldots,k_n} \cdot r_1^{k_1} \cdot \ldots \cdot r_n^{k_n}. \tag{1.26}$$

An equivalent formulation using multisets is:

$$(r_1 + \cdots + r_n)^K = \sum_{\varphi \in \mathcal{M}[K](\{1,\ldots,n\})} \mathbf{(}\varphi\mathbf{)} \cdot \vec{r}^{\,\varphi}. \tag{1.27}$$

There is an 'infinite' version of this result, known as the (Binomial) Series Theorem. It holds more generally than formulated in the first item below, for complex numbers, with adapted meaning of the binomial coefficient, but that's beyond the current scope.

Theorem 1.5.2. *Fix a natural number K.*

1 For a real number $r \in [0, 1)$,

$$\sum_{n \geq 0} \binom{n+K}{K} \cdot r^n = \frac{1}{(1-r)^{K+1}}.$$

2 As a special case,

$$\sum_{n \geq 0} r^n = \frac{1}{1-r}.$$

3 Another consequence is, still for $r \in [0, 1)$,

$$\sum_{n \geq 1} n \cdot r^n = \frac{r}{(1-r)^2}.$$

4 For $r_1, \ldots, r_m \in [0, 1]$ with $\sum_i r_i < 1$,

$$\sum_{n_1,\ldots,n_m \geq 0} \binom{K + \sum_i n_i}{K, n_1, \ldots, n_m} \cdot \prod_i r_i^{n_i} = \frac{1}{(1 - \sum_i r_i)^{K+1}}.$$

Equivalently,

$$\sum_{\varphi\in\mathcal{N}(\{1,\dots,m\})} \binom{K+\|\varphi\|}{K} \cdot (\varphi) \cdot \vec{r}^{\,\varphi} \;=\; \frac{1}{(1-\sum_i r_i)^{K+1}}.$$

Proof. 1 The equation arises as the Taylor series $f(x) = \sum_n \frac{f^{(n)}(0)}{n!} \cdot x^n$ of the function $f(x) = \frac{1}{(1-x)^{K+1}}$. One can show, by induction on n, that the n-th derivative of f is:

$$f^{(n)}(x) \;=\; \frac{(n+K)!}{K!} \cdot \frac{1}{(1-x)^{n+K+1}}.$$

2 The second equation is a special case of the first one, for $K = 0$. There is also a simple direct proof. Define $s_n = r^0 + r^1 + \cdots + r^n$. Then $s_n - r \cdot s_n = 1 - r^{n+1}$, so that $s_n = \frac{1-r^{n+1}}{1-r}$. Hence $s_n \to \frac{1}{1-r}$ as $n \to \infty$.

3 We choose to use the first item, but there are other ways to prove this result, see Exercise 1.5.10.

$$\begin{aligned}\frac{r}{(1-r)^2} \;&=\; r \cdot \sum_{n\geq 0} \binom{n+1}{1} \cdot r^n \qquad \text{by item (1), with } K = 1 \\ &=\; \sum_{n\geq 0} (n+1) \cdot r^{n+1} \;=\; \sum_{n\geq 1} n \cdot r^n.\end{aligned}$$

4 The trick is to turn the multiple sums into a single 'leading' one, in:

$$\begin{aligned}&\sum_{n_1,\dots,n_m\geq 0} \binom{K+\sum_i n_i}{K, n_1, \dots, n_m} \cdot \prod_i r_i^{n_i} \\ &= \sum_{n\geq 0} \sum_{n_1,\dots,n_m,\, \sum_i n_i = n} \binom{K+n}{K, n_1, \dots, n_m} \cdot \prod_i r_i^{n_i} \\ &= \sum_{n\geq 0} \sum_{n_1,\dots,n_m,\, \sum_i n_i = n} \binom{K+n}{K} \cdot \binom{n}{n_1, \dots, n_m} \cdot \prod_i r_i^{n_i} \\ &= \sum_{n\geq 0} \binom{K+n}{K} \cdot \sum_{n_1,\dots,n_m,\, \sum_i n_i = n} \binom{n}{n_1, \dots, n_m} \cdot \prod_i r_i^{n_i} \\ &\overset{(1.26)}{=} \sum_{n\geq 0} \binom{K+n}{K} \cdot (\textstyle\sum_i r_i)^n \\ &= \frac{1}{(1-\sum_i r_i)^{K+1}}, \quad \text{by item (1).}\end{aligned}$$

❑

Exercises

1.5.1 Consider the function $f\colon \{a, b, c\} \to \{0, 1\}$ given by $f(a) = f(b) = 1$ and $f(c) = 0$.

1 Take the natural multiset $\varphi = 1|a\rangle + 3|b\rangle + 1|c\rangle \in \mathcal{N}(\{a, b, c\})$ and compute consecutively:

- $\mathbf{(}\varphi\mathbf{)}$
- $\mathcal{N}(f)(\varphi)$
- $\mathbf{(}\mathcal{N}(f)(\varphi)\mathbf{)}$.

Conclude that $\mathbf{(}\varphi\mathbf{)} \neq \mathbf{(}\mathcal{N}(f)(\varphi)\mathbf{)}$, in general.

2 Now take $\psi = 2|0\rangle + 3|1\rangle \in \mathcal{N}(\{0, 1\})$.

- Compute $\mathbf{(}\psi\mathbf{)}$.
- Show that there are four multisets $\varphi_1, \varphi_2, \varphi_3, \varphi_4 \in \mathcal{N}(\{a, b, c\})$ with $\mathcal{M}(f)(\varphi_i) = \psi$, for each i.
- Check that $\mathbf{(}\psi\mathbf{)} \neq \mathbf{(}\varphi_1\mathbf{)} + \mathbf{(}\varphi_2\mathbf{)} + \mathbf{(}\varphi_3\mathbf{)} + \mathbf{(}\varphi_4\mathbf{)}$.

What is the general formulation now?

1.5.2 Check that:

1 the size map $\| - \| \colon \mathcal{M}(X) \to \mathbb{R}_{\geq 0}$ is a homomorphism of monoids, preserving rescaling — and thus a homomorphism of cones, see Exercise 1.4.9;

2 $\|\mathcal{M}(f)(\varphi)\| = \|\varphi\|$.

1.5.3 Show that for natural multisets $\varphi, \psi \in \mathcal{N}(X)$,

$$\varphi \leq \psi \iff \exists \varphi' \in \mathcal{N}(X).\, \varphi + \varphi' = \psi.$$

1.5.4 Let $\Psi \in \mathcal{N}(\mathcal{N}(X))$ be given with $\psi = \mathit{flat}(\Psi)$. Show that for $\varphi \in \mathit{supp}(\Psi)$ one has $\varphi \leq \psi$ and $\mathit{flat}(\Psi - 1|\varphi\rangle) = \psi - \varphi$.

1.5.5 Let φ be a natural multiset.

1 Show that:

$$\sum_{x \in \mathit{supp}(\varphi)} \frac{\varphi \mathbin{\mathbb{I}\!\circ}}{(\varphi - 1|x\rangle) \mathbin{\mathbb{I}\!\circ}} = \|\varphi\|.$$

$$\begin{aligned}
\sum_{x \in \mathit{supp}(\varphi)} \frac{\varphi \mathbin{\mathbb{I}\!\circ}}{(\varphi - 1|x\rangle) \mathbin{\mathbb{I}\!\circ}} &= \sum_{x \in \mathit{supp}(\varphi)} \frac{\prod_y \varphi(y)!}{(\varphi(x) - 1)! \cdot \prod_{y \neq x} \varphi(y)!} \\
&= \sum_{x \in \mathit{supp}(\varphi)} \frac{\varphi(x)!}{(\varphi(x) - 1)!} \\
&= \sum_{x \in \mathit{supp}(\varphi)} \varphi(x) \\
&= \|\varphi\|.
\end{aligned}$$

2 Derive the recurrence relation (1.25) from this equation.

1.5.6 Prove the multiset formulation (1.27) of the Multinomial Theorem, by induction on K.

1.5.7 Let $K \geq 0$ and let set X have $n \geq 1$ elements. Prove that:

$$\sum_{\varphi \in \mathcal{N}[K](X)} \binom{\varphi}{} = n^K.$$

This generalises the well known sum-formula for binomial coefficients: $\sum_{0 \leq k \leq K} \binom{K}{k} = 2^K$, for $n = 2$.
Hint: Use $n = 1 + \cdots + 1$ in (1.27).

1.5.8 Let φ be a natural multiset. Show that:

1. $\binom{\varphi}{} = 1 \iff \mathit{supp}(\varphi)$ is a singleton;
2. $\binom{\varphi}{} = \|\varphi\|! \iff \forall x \in \mathit{supp}(\varphi).\, \varphi(x) = 1 \iff \varphi$ consists of singletons.

1.5.9 Let $n \geq 1$ and $r \in (0, 1)$. Show that:

$$\sum_{k \geq n} \binom{k-1}{n-1} \cdot r^n \cdot (1-r)^{k-n} = 1.$$

1.5.10 Elaborate the details of the following two (alternative) proofs of the equation in Theorem 1.5.2 (3).

1. Use the derivate $\frac{d}{dr}$ on both sides of Theorem 1.5.2 (2).
2. Write $s \coloneqq \sum_{n \geq 1} n \cdot r^n$ and exploit the following recursive equation.

$$\begin{aligned} s &= r + 2r^2 + 3r^3 + 4r^4 + \cdots \\ &= r + (1+1)r^2 + (1+2)r^3 + (1+3)r^4 + \cdots \\ &= (r + r^2 + r^3 + r^4 + \cdots) + r \cdot (r + 2r^2 + 3r^3 + \cdots) \\ &= \frac{r}{1-r} + r \cdot s, \qquad \text{by Theorem 1.5.2 (2).} \end{aligned}$$

1.5.11 In the proof of Theorem 1.5.2 we have used Taylor's formula for a single-variable function. For multi-variable functions we can use multisets for a compact, 'multi-index' formulation. For an an n-ary function $f(x_1, \ldots, x_n)$ and a natural multiset $\varphi \in \mathcal{N}(\{1, \ldots, n\})$ write:

$$\partial^\varphi f \coloneqq (\partial x_1)^{\varphi(1)} \cdots (\partial x_n)^{\varphi(n)} f.$$

Check that Taylor's expansion formula (around $0 \in \mathbb{R}^n$) then becomes:

$$f(\vec{x}) = \sum_{\varphi \in \mathcal{N}(\{1,\ldots,n\})} \frac{(\partial^\varphi f)(0)}{\varphi!} \cdot \vec{x}^{\,\varphi}.$$

1.6 Binomial coefficients of multisets

Binomial coefficients $\binom{n}{k}$ for numbers $n \geq k$ are a standard tool in many areas of mathematics, see Subsection 1.3.4 for the definition and the most basic properties. Here we extend binomial coefficients from numbers to natural multisets: we define $\binom{\psi}{\varphi}$ for natural multisets ψ, φ with $\psi \geq \varphi$. In the next section we shall also look into the less familiar 'multichoose' coefficients $\binom{\binom{n}{m}}{}$ and their extension to multisets $\binom{\binom{\psi}{\varphi}}{}$. The coefficients $\binom{\psi}{\varphi}$ and $\binom{\binom{\psi}{\varphi}}{}$ will play an important role in (multivariate) hypergeometric and Pólya distributions.

Definition 1.6.1. For natural multisets $\varphi, \psi \in \mathcal{N}(X)$ with $\varphi \leq \psi$, the *multiset binomial* is defined as:

$$\begin{aligned}\binom{\psi}{\varphi} &:= \frac{\psi!}{\varphi! \cdot (\psi - \varphi)!} \\ &= \frac{\prod_x \psi(x)!}{(\prod_x \varphi(x)!) \cdot (\prod_x (\psi(x) - \varphi(x))!)} = \prod_{x \in supp(\psi)} \binom{\psi(x)}{\varphi(x)}.\end{aligned} \tag{1.28}$$

For instance:

$$\binom{3|R\rangle + 2|B\rangle}{2|R\rangle + 1|B\rangle} = \frac{3! \cdot 2!}{(2! \cdot 1!) \cdot (1! \cdot 1!)} = 6 = 3 \cdot 2 = \binom{3}{2} \cdot \binom{2}{1}.$$

This describes the number of possible ways of drawing $2|R\rangle + 1|B\rangle$ from an urn $3|R\rangle + 2|B\rangle$.

The following result is a generalisation of Vandermonde's formula.

Lemma 1.6.2. *Let $\psi \in \mathcal{N}(X)$ be a multiset of size $L = \|\psi\|$, with a number $K \leq L$. Then:*

$$\sum_{\varphi \leq_K \psi} \binom{\psi}{\varphi} = \binom{L}{K} \qquad \text{so that} \qquad \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L}{K}} = 1.$$

These fractions adding up to one will form the probabilities of the so-called hypergeometric distributions, see Example 2.1.1 (3) and Section 3.4 later on.

Proof. We use induction on the number of elements in $supp(\psi)$. We go through some initial values explicitly. If the number of elements is 0, then $\psi = \mathbf{0}$ and so $L = 0 = K$ and $\varphi \leq_K \psi$ means $\varphi = \mathbf{0}$, so that the result holds. Similarly, if $supp(\psi)$ is a singleton, say $\{x\}$, then $L = \psi(x)$. For $K \leq L$ and $\varphi \leq_K \psi$ we get $supp(\varphi) = \{x\}$ and $K = \varphi(x)$. The result then obviously holds.

The case where $supp(\psi) = \{x, y\}$ captures the ordinary form of Vandermonde's formula. We reformulate it for numbers $B, G \in \mathbb{N}$ and $K \leq B + G$.

Then:

$$\binom{B+G}{K} = \sum_{b\leq B,\, g\leq G,\, b+g=K} \binom{B}{b}\cdot\binom{G}{g}. \tag{1.29}$$

Intuitively: if you select K children out of B boys and G girls, the number of options is given by the sum over the options for $b \leq B$ boys times the options for $g \leq G$ girls, with $b + g = K$.

The equation (1.29) can be proven by induction on G. When $G = 0$ both sides amount to $\binom{B}{K}$ so we quickly proceed to the induction step. The case $K = 0$ is trivial, so we may assume $K > 0$.

$$\begin{aligned}
&\sum_{b\leq B,\, g\leq G+1,\, b+g=K} \binom{B}{b}\cdot\binom{G+1}{g} \\
&= \binom{B}{K}\cdot\binom{G+1}{0} + \binom{B}{K-1}\cdot\binom{G+1}{1} + \cdots + \binom{B}{K-G}\cdot\binom{G+1}{G} + \binom{B}{K-G-1}\cdot\binom{G+1}{G+1} \\
&= \binom{B}{K}\cdot\binom{G}{0} + \binom{B}{K-1}\cdot\binom{G}{1} + \binom{B}{K-1}\cdot\binom{G}{0} \\
&\qquad + \cdots + \binom{B}{K-G}\cdot\binom{G}{G} + \binom{B}{K-G}\cdot\binom{G}{G-1} + \binom{B}{K-G-1}\cdot\binom{G}{G} \\
&= \sum_{b\leq B,\, g\leq G,\, b+g=K} \binom{B}{b}\cdot\binom{G}{g} + \sum_{b\leq B,\, g\leq G,\, b+g=K-1} \binom{B}{b}\cdot\binom{G}{g} \\
&\overset{\text{(IH)}}{=} \binom{B+G}{K} + \binom{B+G}{K-1} \\
&\overset{(1.14)}{=} \binom{B+G+1}{K}.
\end{aligned}$$

For the induction step, let $\mathit{supp}(\psi) = \{x_1, \ldots, x_n, y\}$, for $n \geq 2$. Writing $\ell = \psi(y)$, $L' = L - \ell$ and $\psi' = \psi - \ell|y\rangle \in \mathcal{N}[L'](X)$ gives:

$$\begin{aligned}
\sum_{\varphi\leq_K \psi} \binom{\psi}{\varphi} = \sum_{\varphi\leq_K \psi} \prod\nolimits_x \binom{\psi(x)}{\varphi(x)} &= \sum_{n\leq \ell} \sum_{\varphi\leq_{K-n}\psi'} \binom{\ell}{n}\cdot\prod\nolimits_i \binom{\psi'(x_i)}{\varphi(x_i)} \\
&\overset{\text{(IH)}}{=} \sum_{n\leq \ell,\, K-n\leq L-\ell} \binom{\ell}{n}\cdot\binom{L-\ell}{K-n} \overset{(1.29)}{=} \binom{L}{K}.
\end{aligned}$$

❑

For a multiset φ we have already used the 'support' definition $\mathit{supp}(\varphi) = \{x \mid \varphi(x) \neq 0\}$. This yields a map $\mathit{supp}\colon \mathcal{M}(X) \to \mathcal{P}_{\mathit{fin}}(X)$, which is well-behaved, in the sense that it is natural and preserves the monoid structures on $\mathcal{M}(X)$ and $\mathcal{P}_{\mathit{fin}}(X)$.

We have also seen a support map from lists $\mathcal{L}$ to finite powerset $\mathcal{P}_{\mathit{fin}}$. This support map factorises through multisets, as described with a new function *acc* in the following triangle.

$$\begin{array}{ccc}
\mathcal{L}(X) & \xrightarrow{\quad\mathit{supp}\quad} & \mathcal{P}_{\mathit{fin}}(X) \\
& {\scriptstyle \mathit{acc}}\searrow \quad \nearrow{\scriptstyle \mathit{supp}} & \\
& \mathcal{N}(X) &
\end{array} \tag{1.30}$$

The 'accumulator' map $\mathit{acc}\colon \mathcal{L}(X) \to \mathcal{N}(X)$ counts (accumulates) how many

times an element occurs in a list, while ignoring the order of occurrences. Thus, for a list $\ell \in \mathcal{L}(X)$,

$$acc(\ell)(x) \coloneqq n \quad \text{if } x \text{ occurs } n \text{ times in the list } \alpha. \tag{1.31}$$

Alternatively,

$$acc(x_1, \ldots, x_n) \;=\; 1|x_1\rangle + \cdots + 1|x_n\rangle.$$

Or, more concretely, in an example:

$$acc(a, b, a, b, c, b, b) \;=\; 2|a\rangle + 4|b\rangle + 1|c\rangle.$$

The above diagram (1.30) expresses an earlier informal statement, namely that multisets are somehow in between lists and subsets.

The starting point in this section is the question: how many (ordered) sequences of coloured balls give rise to a specific urn? More technically, given a natural multiset φ, how many lists ℓ statisfy $acc(\ell) = \varphi$? In yet another form, what is the size $|acc^{-1}(\varphi)|$ of the inverse image?

As described in the beginning of this section, one aim is to relate the multiset coefficient $\mathbf{(}\varphi\mathbf{)}$ of a multiset φ to the number of lists that accumlate to φ, as defined in (1.31). Here we use a K-ary version of accumulation, for $K \in \mathbb{N}$, restricted to K-many elements. It then becomes a mapping:

$$X^K \xrightarrow{\;acc[K]\;} \mathcal{N}[K](X). \tag{1.32}$$

The parameter K will often be omitted when it is clear from the context. We are now ready for a basic combinatorial / counting result. It involves the multiset coefficient from Definition 1.5.1 (5).

Proposition 1.6.3. *For $\varphi \in \mathcal{N}[K](X)$ one has:*

$$\mathbf{(}\varphi\mathbf{)} \;=\; \left|\, acc^{-1}(\varphi) \,\right| \;=\; \textit{the number of lists } \ell \in X^K \textit{ with } acc(\ell) = \varphi.$$

Proof. We use induction on the number of elements of the support $supp(\varphi)$ of the multiset φ. If this number is 0, then $\varphi = \mathbf{0}$, with $\mathbf{(0)} = 1$. And indeed, there is precisely one list ℓ with $acc(\ell) = \mathbf{0}$, namely the empty list [].

Next suppose that $supp(\varphi) = \{x_1, \ldots, x_n, x_{n+1}\}$. Take $m \coloneqq \varphi(x_{n+1})$ and $\varphi' \coloneqq \varphi - m|x_{n+1}\rangle$ so that $\|\varphi'\| = K - m$ and $supp(\varphi') = \{x_1, \ldots, x_n\}$. By the induction hypothesis there are $\mathbf{(}\varphi'\mathbf{)}$-many lists $\ell' \in X^{K-m}$ with $acc(\ell') = \varphi'$. Each such list ℓ' can be extended to a list ℓ with $acc(\ell) = \varphi$ by m times adding x_{n+1} to ℓ'. How many such additions are there? It is not hard to see that this number of

additions is $\binom{K}{m}$. Thus:

$$\begin{aligned}\left|\mathit{acc}^{-1}(\varphi)\right| &= \binom{K}{m} \cdot \mathbf{(}\varphi'\mathbf{)} \\ &= \frac{K!}{m!(K-m)!} \cdot \frac{(K-m)!}{\prod_{i\leq n} \varphi'(x_i)!} \\ &= \frac{K!}{\prod_{i\leq n+1} \varphi(x_i)!} \qquad \text{since } m = \varphi(x_{n+1}) \text{ and } \varphi'(x_i) = \varphi(x_i) \\ &= \mathbf{(}\varphi\mathbf{)} \end{aligned}$$

❑

We conclude this part with a remarkable fact, combining multinomial coefficients with the flatten operation for multisets, and also with accumulation. We start with an illustration for which we recall from Subsection 1.4.2 that the flatten operation has type $\mathit{flat}\colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$. When we restrict it to natural multisets and include sizes $K, L \in \mathbb{N}$ it takes the form:

$$\mathcal{M}[L]\Big(\mathcal{M}[K](X)\Big) \xrightarrow{\mathit{flat}} \mathcal{M}[L\cdot K](X) \tag{1.33}$$

Now suppose a natural multiset is given:

$$\psi \;=\; 3|a\rangle + 1|b\rangle + 2|c\rangle \;\in\; \mathcal{M}[2\cdot 3](X) \qquad \text{for} \qquad \begin{cases} K \;=\; 3,\; L \;=\; 2 \\ X \;=\; \{a,b,c\}. \end{cases}$$

It's multinomial coefficient is $\mathbf{(}\psi\mathbf{)} = \frac{6!}{3!\cdot 1!\cdot 2!} = 60$.

We now ask ourselves the question: which $\Psi \in \mathcal{N}[L](\mathcal{N}[K](X))$ satisfy $\mathit{flat}(\Psi) = \psi$? A little reflection shows that there are three such Ψ, namely:

$$\begin{aligned}\Psi_1 &= 1\Big|2|a\rangle + 1|c\rangle\Big\rangle + 1\Big|1|a\rangle + 1|b\rangle + 1|c\rangle\Big\rangle \\ \Psi_2 &= 1\Big|2|a\rangle + 1|b\rangle\Big\rangle + 1\Big|1|a\rangle + 2|c\rangle\Big\rangle \\ \Psi_3 &= 1\Big|3|a\rangle\Big\rangle + 1\Big|1|b\rangle + 2|c\rangle\Big\rangle.\end{aligned}$$

Next we are going to compute the expression $\mathbf{(}\Psi\mathbf{)} \cdot \prod_\varphi \mathbf{(}\varphi\mathbf{)}^{\Psi(\varphi)}$ in each case.

$$\begin{aligned}\mathbf{(}\Psi_1\mathbf{)} \cdot \textstyle\prod_\varphi \mathbf{(}\varphi\mathbf{)}^{\Psi_1(\varphi)} &= 2 \cdot 3^1 \cdot 6^1 \;=\; 36 \\ \mathbf{(}\Psi_2\mathbf{)} \cdot \textstyle\prod_\varphi \mathbf{(}\varphi\mathbf{)}^{\Psi_2(\varphi)} &= 2 \cdot 3^1 \cdot 3^1 \;=\; 18 \\ \mathbf{(}\Psi_3\mathbf{)} \cdot \textstyle\prod_\varphi \mathbf{(}\varphi\mathbf{)}^{\Psi_3(\varphi)} &= 2 \cdot 1^1 \cdot 3^1 \;=\; 6\end{aligned}$$

We then find that the sum of these outcomes equals the coefficient of ψ:

$$\begin{aligned}\mathbf{(}\psi\mathbf{)} &= 60 \\ &= \Big(\mathbf{(}\Psi_1\mathbf{)}\cdot\textstyle\prod_\varphi \mathbf{(}\varphi\mathbf{)}^{\Psi_1(\varphi)}\Big) + \Big(\mathbf{(}\Psi_2\mathbf{)}\cdot\textstyle\prod_\varphi \mathbf{(}\varphi\mathbf{)}^{\Psi_2(\varphi)}\Big) + \Big(\mathbf{(}\Psi_3\mathbf{)}\cdot\textstyle\prod_\varphi \mathbf{(}\varphi\mathbf{)}^{\Psi_3(\varphi)}\Big).\end{aligned}$$

The general result is formulated in Theorem 1.6.5 below. For its proof we need an intermediate step.

Lemma 1.6.4. *Let $\psi \in \mathcal{N}[L](X)$ be a natural multiset of size $L \in \mathbb{N}$, and let $K \leq L$.*

1 For $\varphi \leq_K \psi$ one has:

$$\frac{\mathbf{(}\varphi\mathbf{)} \cdot \mathbf{(}\psi - \varphi\mathbf{)}}{\mathbf{(}\psi\mathbf{)}} = \frac{\binom{\psi}{\varphi}}{\binom{L}{K}}.$$

2 Now:

$$\mathbf{(}\psi\mathbf{)} = \sum_{\varphi \leq_K \psi} \mathbf{(}\varphi\mathbf{)} \cdot \mathbf{(}\psi - \varphi\mathbf{)}.$$

The earlier equation (1.25) *is a special case, for $K = 1$.*

Proof. 1 Because:

$$\begin{aligned} \frac{\mathbf{(}\varphi\mathbf{)} \cdot \mathbf{(}\psi - \varphi\mathbf{)}}{\mathbf{(}\psi\mathbf{)}} &= \frac{K!}{\varphi\mathinner{!}_\circ} \cdot \frac{(L-K)!}{(\psi - \varphi)\mathinner{!}_\circ} \cdot \frac{\psi\mathinner{!}_\circ}{L!} \\ &= \frac{K! \cdot (L-K)!}{L!} \cdot \frac{\psi\mathinner{!}_\circ}{\varphi\mathinner{!}_\circ \cdot (\psi - \varphi)\mathinner{!}_\circ} = \frac{\binom{\psi}{\varphi}}{\binom{L}{K}}. \end{aligned}$$

2 By the previous item and Vandermonde's formula from Lemma 1.6.2:

$$\sum_{\varphi \leq_K \psi} \mathbf{(}\varphi\mathbf{)} \cdot \mathbf{(}\psi - \varphi\mathbf{)} = \mathbf{(}\psi\mathbf{)} \cdot \frac{\sum_{\varphi \leq_K \psi} \binom{\psi}{\varphi}}{\binom{L}{K}} = \mathbf{(}\psi\mathbf{)} \cdot \frac{\binom{L}{K}}{\binom{L}{K}} = \mathbf{(}\psi\mathbf{)}. \qquad \text{❑}$$

Theorem 1.6.5. *Each $\psi \in \mathcal{N}[L \cdot K](X)$, for numbers $L, K \geq 1$, satisfies:*

$$\mathbf{(}\psi\mathbf{)} = \sum_{\substack{\Psi \in \mathcal{N}[L](\mathcal{N}[K](X)) \\ \mathit{flat}(\Psi) = \psi}} \mathbf{(}\Psi\mathbf{)} \cdot \prod\nolimits_{\varphi} \mathbf{(}\varphi\mathbf{)}^{\Psi(\varphi)}.$$

This equation turns out to be essential for proving that multinomial distributions are suitably closed under composition, see Theorem 3.3.6 in Chapter 3.

Proof. Since $\psi \in \mathcal{N}[L \cdot K](X)$ we can apply Lemma 1.6.4 (2) L-many times, giving:

$$\begin{aligned} \mathbf{(}\psi\mathbf{)} &= \sum_{\varphi_1 \leq_K \psi} \sum_{\varphi_2 \leq_K \psi - \varphi_1} \cdots \sum_{\varphi_L \leq_K \psi - \varphi_1 - \cdots - \varphi_{L-1}} \mathbf{(}\varphi_1\mathbf{)} \cdot \mathbf{(}\varphi_2\mathbf{)} \cdot \ldots \cdot \mathbf{(}\varphi_L\mathbf{)} \\ &= \sum_{\substack{\varphi_1, \ldots, \varphi_L \leq_K \psi \\ \varphi_1 + \cdots + \varphi_L = \psi}} \prod\nolimits_i \mathbf{(}\varphi_i\mathbf{)}. \end{aligned}$$

We can accumulate the sequence of multisets $\varphi_1, \ldots, \varphi_L \in \mathcal{N}[K]$ into a multiset of multisets:

$$\Psi \coloneqq acc(\varphi_1, \ldots, \varphi_L) = 1|\varphi_1\rangle + \cdots + 1|\varphi_L\rangle \in \mathcal{N}[L](\mathcal{N}[K](X)).$$

The flatten map preserves sums of multisets, see Exercise 1.4.7, and thus maps Ψ to ψ, via the flatten-unit law of Lemma 1.4.3.

$$\begin{aligned} flat(\Psi) &= flat\Big(1|\varphi_1\rangle + \cdots + 1|\varphi_L\rangle\Big) \\ &= flat(1|\varphi_1\rangle) + \cdots + flat(1|\varphi_L\rangle) \\ &= flat(unit(\varphi_1)) + \cdots + flat(unit(\varphi_L)) = \varphi_1 + \cdots + \varphi_L = \psi. \end{aligned}$$

By summing over the accumulation $\Psi \in \mathcal{N}[L](\mathcal{N}[K](X))$ with $flat(\Psi) = \psi$, instead of over the sequences $\varphi_1, \ldots, \varphi_L \leq_K \psi$ with $\sum_i \varphi_i = \psi$ we have that into account that $\mathbf{(}\Psi\mathbf{)}$-many sequences accumulate to the same Ψ, see Proposition 1.6.3. This explains the factor $\mathbf{(}\Psi\mathbf{)}$ in the above theorem. ❑

Exercises

1.6.1 Generalise the familiar equation $\sum_{0\leq k\leq K} \binom{K}{k} = 2^K$ to:

$$\sum_{\varphi\leq\psi} \binom{\psi}{\varphi} = 2^{\|\psi\|}.$$

1.6.2 Show that $\|acc(\ell)\| = \|\ell\|$, using the length $\|\ell\|$ of a list ℓ from Exercise 1.2.3.

1.6.3 Let $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](X)$ be given.

1 Prove that:

$$\mathbf{(}\varphi + \psi\mathbf{)} = \frac{\binom{K+L}{K}}{\binom{\varphi+\psi}{\varphi}} \cdot \mathbf{(}\varphi\mathbf{)} \cdot \mathbf{(}\psi\mathbf{)}.$$

2 Now assume that φ, ψ have disjoint supports, that is, $supp(\varphi) \cap supp(\psi) = \emptyset$. Show that now:

$$\mathbf{(}\varphi + \psi\mathbf{)} = \binom{K+L}{K} \cdot \mathbf{(}\varphi\mathbf{)} \cdot \mathbf{(}\psi\mathbf{)},$$

1.6.4 Consider the K-ary accumulator function (1.32), for $K > 0$. Check that *acc* is permutation-invariant, in the sense that for each permutation π of $\{1, \ldots, K\}$ one has:

$$acc(x_1, \ldots, x_K) = acc(x_{\pi(1)}, \ldots, x_{\pi(K)}).$$

1.6.5 Check that the accumulator map $acc\colon \mathcal{L}(X) \to \mathcal{M}(X)$ is a homomorphism of monoids. Do the same for the support map $supp\colon \mathcal{M}(X) \to \mathcal{P}_{fin}(X)$.

1.6.6 Prove that the accumulator and support maps $acc\colon \mathcal{L}(X) \to \mathcal{M}(X)$ and $supp\colon \mathcal{M}(X) \to \mathcal{P}_{fin}(X)$ are natural: for an arbitrary function $f\colon X \to Y$ both rectangles below commute.

$$\begin{array}{ccccc} \mathcal{L}(X) & \xrightarrow{acc} & \mathcal{M}(X) & \xrightarrow{supp} & \mathcal{P}_{fin}(X) \\ {\scriptstyle \mathcal{L}(f)}\downarrow & & \downarrow{\scriptstyle \mathcal{M}(f)} & & \downarrow{\scriptstyle \mathcal{P}_{fin}(f)} \\ \mathcal{L}(Y) & \xrightarrow[acc]{} & \mathcal{M}(Y) & \xrightarrow[supp]{} & \mathcal{P}_{fin}(Y) \end{array}$$

1.6.7 Convince yourself that the following composite

$$\mathcal{N}[K](X)^L \xrightarrow{acc} \mathcal{N}[L]\Big(\mathcal{N}[K](X)\Big) \xrightarrow{flat} \mathcal{N}[L{\cdot}K](X)$$

is the L-fold sum of multisets — see (1.33) for the fixed size flatten operation.

1.6.8 In analogy with the powerset operator, with type $\mathcal{P}\colon \mathcal{P}(X) \to \mathcal{P}(\mathcal{P}(X))$, a *powerbag* operator $\mathsf{PB}\colon \mathcal{N}(X) \to \mathcal{N}(\mathcal{N}(X))$ is introduced in [114] (see also [35]). It can be defined as:

$$\mathsf{PB}(\psi) \coloneqq \sum_{\varphi \leq \psi} \binom{\psi}{\varphi} \big|\varphi\rangle.$$

1 Take $X = \{a, b\}$ and show that:

$$\begin{aligned} &\mathsf{PB}(1|a\rangle + 3|b\rangle) \\ &= 1\big|\mathbf{0}\big\rangle + 1\big|1|a\rangle\big\rangle + 3\big|1|b\rangle\big\rangle + 3\big|1|a\rangle + 1|b\rangle\big\rangle + 3\big|2|b\rangle\big\rangle \\ &\qquad + 3\big|1|a\rangle + 2|b\rangle\big\rangle + 1\big|3|b\rangle\big\rangle + 1\big|1|a\rangle + 3|b\rangle\big\rangle. \end{aligned}$$

2 Check that one can compute the powerbag of ψ as follows. Take a list of elements that accumulate to ψ, such as $[a, b, b, b]$ in the previous item. Take the accumulation of all subsequences.

1.6.9 For $N \geq 2$ many natural multisets $\varphi_1, \ldots, \varphi_N \in \mathcal{N}(X)$, with $\psi \coloneqq \sum_i \varphi_i$, define a multinomial coefficient of multisets as:

$$\binom{\psi}{\varphi_1, \ldots, \varphi_N} \coloneqq \frac{\psi!}{\varphi_1! \cdot \ldots \cdot \varphi_N!}.$$

1 Check that for $N \geq 3$, in analogy with Exercise 1.3.6,

$$\binom{\psi}{\varphi_1, \ldots, \varphi_N} = \binom{\psi}{\varphi_1} \cdot \binom{\psi - \varphi_1}{\varphi_2, \ldots, \varphi_N}$$

2 For $K_1, \ldots, K_N \in \mathbb{N}$ write $K = \sum_i K_i$ and assume that $\psi \in \mathcal{N}[K](X)$ is given. Show that:

$$\sum_{\substack{\varphi_1 \leq_{K_1} \psi, \ldots, \varphi_N \leq_{K_N} \psi, \\ \sum_i \varphi_i = \psi}} \binom{\psi}{\varphi_1, \ldots, \varphi_N} = \binom{K}{K_1, \ldots, K_N}.$$

1.7 Multichoose coefficents

We proceed with another counting challenge. Let X be a finite set, say with n elements. How many elements are there then in $\mathcal{N}[K](X)$? That is, how many multisets of size K are there over n elements? This is sometimes formulated informally as: how many ways are there to divide K balls over n urns? It is the multiset-analogue of Lemma 1.3.7, where the number of *subsets* of size K of an n-element set is identified as $\binom{n}{K}$. Below we show that the answer for multisets is given by the *multiset number* or *multichoose number* $\left(\!\binom{n}{K}\!\right)$, see *e.g.* [145]. We introduce this new number below, and immediately extend its standard definition from numbers to multisets, in analogy with the extension of (ordinary) binomials to multisets in Definition 1.5.1. We first need some preparations before coming to the multiset counting result in Proposition 1.7.4.

Definition 1.7.1. 1 For numbers $n \geq 1$ and $K \geq 0$, put:

$$\left(\!\binom{n}{K}\!\right) := \binom{n+K-1}{K} = \binom{n+K-1}{n-1} = \frac{(n+K-1)!}{K! \cdot (n-1)!} = \frac{n}{n+K} \cdot \binom{n+K}{n}. \tag{1.34}$$

This $\left(\!\binom{n}{K}\!\right)$ is sometimes called *multichoose*.

2 Let ψ, φ be natural multisets on the same space, where ψ is non-empty and $\mathit{supp}(\varphi) \subseteq \mathit{supp}(\psi)$.

$$\left(\!\binom{\psi}{\varphi}\!\right) := \prod_{x \in \mathit{supp}(\psi)} \left(\!\binom{\psi(x)}{\varphi(x)}\!\right) = \prod_{x \in \mathit{supp}(\psi)} \binom{\psi(x) + \varphi(x) - 1}{\varphi(x)}. \tag{1.35}$$

Consider the set $\mathcal{N}[3](\{a, b, c\})$ of multisets of size 3 over $\{a, b, c\}$. It has $\left(\!\binom{3}{3}\!\right) = \binom{5}{3} = \frac{4 \cdot 5}{2} = 10$ elements, namely:

$$\begin{aligned} &3|a\rangle, \quad 3|b\rangle, \quad 3|c\rangle, \quad 2|a\rangle + 1|b\rangle, \quad 2|a\rangle + 1|c\rangle, \quad 1|a\rangle + 2|b\rangle, \\ &2|b\rangle + 1|c\rangle, \quad 1|a\rangle + 2|c\rangle, \quad 1|b\rangle + 2|c\rangle, \quad 1|a\rangle + 1|b\rangle + 1|c\rangle. \end{aligned}$$

Our aim is to prove that the multiset number $\left(\!\binom{n}{K}\!\right)$ is the total number of multisets of size K over a non-empty n-element set. We use the following multichoose analogue of Vandermonde's (binary) formula (1.29).

Lemma 1.7.2. *Fix $B \geq 1$ and $G \geq 1$. For all K on has:*

$$\left(\!\binom{B+G}{K}\!\right) = \sum_{0\leq k\leq K} \left(\!\binom{B}{k}\!\right) \cdot \left(\!\binom{G}{K-k}\!\right). \tag{1.36}$$

In particular:

$$\binom{B+K}{K} = \left(\!\binom{B+1}{K}\!\right) = \sum_{0\leq k\leq K} \left(\!\binom{B}{k}\!\right) = \sum_{0\leq k\leq K} \left(\!\binom{B}{K-k}\!\right). \tag{1.37}$$

Proof. The second equation (1.37) easily follows from the first one by taking $G = 1$ and using that $\left(\!\binom{1}{n}\!\right) = \binom{n}{n} = 1$.

We shall make frequent use of the following equation, whose proof is left to the reader (in Exercise 1.7.5) below.

$$\left(\!\binom{n}{K+1}\!\right) + \left(\!\binom{n+1}{K}\!\right) = \left(\!\binom{n+1}{K+1}\!\right). \tag{$*$}$$

We shall prove the first equation (1.36) in the lemma by induction on $B \geq 1$. In both the base case $B = 1$ and the induction step we shall use induction on K. We shall try to keep the structure clear by using nested bullets.

- We first prove Equation (1.36) for $B = 1$, by induction on K.
  - When $K = 0$ both sides in (1.36) are equal to 1.
  - Assume Equation (1.36) holds for K (and $B = 1$).

$$\begin{aligned}\sum_{0\leq k\leq K+1} \left(\!\binom{1}{k}\!\right) \cdot \left(\!\binom{G}{(K+1)-k}\!\right) = \sum_{0\leq k\leq K+1} \left(\!\binom{G}{K-(k-1)}\!\right) &= \left(\!\binom{G}{K+1}\!\right) + \sum_{0\leq \ell\leq K} \left(\!\binom{G}{K-\ell}\!\right) \\ &\overset{\text{(IH)}}{=} \left(\!\binom{G}{K+1}\!\right) + \left(\!\binom{G+1}{K}\!\right) \\ &\overset{(*)}{=} \left(\!\binom{G+1}{K+1}\!\right).\end{aligned}$$

- Now assume Equation (1.36) holds for B (for all G, K). In order to show that it then also holds for $B + 1$ we use induction on K.
  - When $K = 0$ both sides in (1.36) are equal to 1.

– Now assume that Equation (1.36) holds for K, and for B. Then:

$$\begin{aligned}
&\sum_{0\leq k\leq K+1} \left(\!\binom{B+1}{k}\!\right)\cdot\left(\!\binom{G}{(K+1)-k}\!\right)\\
&= \left(\!\binom{G}{K+1}\!\right) + \sum_{0\leq k\leq K} \left(\!\binom{B+1}{k+1}\!\right)\cdot\left(\!\binom{G}{K-k}\!\right)\\
&\overset{(*)}{=} \left(\!\binom{G}{K+1}\!\right) + \sum_{0\leq k\leq K} \left[\left(\!\binom{B}{k+1}\!\right) + \left(\!\binom{B+1}{k}\!\right)\right]\cdot\left(\!\binom{G}{K-k}\!\right)\\
&= \left(\!\binom{G}{K+1}\!\right) + \sum_{0\leq k\leq K} \left(\!\binom{B}{k+1}\!\right)\cdot\left(\!\binom{G}{K-k}\!\right) + \sum_{0\leq k\leq K} \left(\!\binom{B+1}{k}\!\right)\cdot\left(\!\binom{G}{K-k}\!\right)\\
&\overset{\text{(IH, K)}}{=} \sum_{0\leq k\leq K+1} \left(\!\binom{B}{k}\!\right)\cdot\left(\!\binom{G}{(K+1)-k}\!\right) + \left(\!\binom{(B+1)+G}{K}\!\right)\\
&\overset{\text{(IH, B)}}{=} \left(\!\binom{B+G}{K+1}\!\right) + \left(\!\binom{(B+1)+G}{K}\!\right)\\
&\overset{(*)}{=} \left(\!\binom{(B+1)+G}{K+1}\!\right).
\end{aligned}$$

❑

We now get the double-bracket analogue of Lemma 1.6.2.

Proposition 1.7.3. *Let ψ be a non-empty natural multiset. Write $X = supp(\psi)$ and $L = \|\psi\|$. Then, for each $K \in \mathbb{N}$,*

$$\sum_{\varphi\in\mathcal{N}[K](X)} \left(\!\binom{\psi}{\varphi}\!\right) = \left(\!\binom{L}{K}\!\right) \qquad \textit{so} \qquad \sum_{\varphi\in\mathcal{N}[K](X)} \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} = 1.$$

The fractions in this equation will show up later in so-called Pólya distributions, see Example 2.1.1 (4) and Section 3.4. These fractions capture the probability of drawing a multiset φ from an urn ψ when for each drawn ball an extra ball is added to the urn (of the same colour).

Proof. We use induction on the number of elements in the support X of ψ, like in the proof of Lemma 1.6.2. By assumption X cannot be empty, so the induction starts when X is a singleton, say $X = \{x\}$. But then $\psi(x) = \|\psi\| = L$ and $\varphi(x) = \|\varphi\| = K$, so the result obviously holds.

Now let $supp(\psi) = X \cup \{y\}$ where $y \notin X$ and X is not empty. Write:

$$L = \|\psi\| \qquad \ell = \psi(y) > 0 \qquad \psi' = \psi - \ell|y\rangle \qquad L' = L - \ell > 0.$$

By construction $X = supp(\psi')$ and $L' = \|\psi'\|$. Now:

$$\begin{aligned}
\sum_{\varphi\in\mathcal{N}[K](X\cup\{y\})} \left(\!\binom{\psi}{\varphi}\!\right) &\overset{(1.35)}{=} \sum_{\varphi\in\mathcal{N}[K](X\cup\{y\})} \prod_{x\in X\cup\{y\}} \left(\!\binom{\psi(x)}{\varphi(x)}\!\right) \\
&= \sum_{0\leq k\leq K} \sum_{\varphi\in\mathcal{N}[K-k](X)} \left(\!\binom{\psi(y)}{k}\!\right) \cdot \prod_{x\in X} \left(\!\binom{\psi(x)}{\varphi(x)}\!\right) \\
&= \sum_{0\leq k\leq K} \left(\!\binom{\ell}{k}\!\right) \cdot \sum_{\varphi\in\mathcal{N}[K-k](X)} \left(\!\binom{\psi}{\varphi}\!\right) \\
&\overset{(IH)}{=} \sum_{0\leq k\leq K} \left(\!\binom{\ell}{k}\!\right) \cdot \left(\!\binom{L'}{K-k}\!\right) \\
&\overset{(1.36)}{=} \left(\!\binom{\ell+L'}{K}\!\right) = \left(\!\binom{L}{K}\!\right). \qquad \text{❑}
\end{aligned}$$

We finally come to our multiset counting result. It is the multiset-analogue of Proposition 1.3.7 (1) for subsets.

Proposition 1.7.4. *Let X be a non-empty set with $n \geq 1$ elements. The number of natural multisets of size K over X is $\left(\!\binom{n}{K}\!\right)$, that is:*

$$\left|\mathcal{N}[K](X)\right| = \left(\!\binom{n}{K}\!\right) = \binom{n+K-1}{K}.$$

Proof. The statement holds for $K = 0$ since there is precisely $1 = \binom{n-1}{0} = \left(\!\binom{n}{0}\!\right)$ multiset set of 0, namely the empty multiset $\mathbf{0}$. Hence we may assume $K \geq 1$, so that Lemma 1.7.2 can be used.

We proceed by induction on $n \geq 1$. For $n = 1$ the statement holds since there is only $1 = \binom{K}{K} = \left(\!\binom{1}{K}\!\right)$ multiset of size K over $1 = \{0\}$, namely $K|0\rangle$.

The induction step works as follows. Let X have n elements, and $y \notin X$. For a multiset $\varphi \in \mathcal{N}[K](X \cup \{y\})$ there are $K + 1$ possible multiplicities $\varphi(n)$. If $\varphi(n) = k$, then the number possibilities for $\varphi(0), \ldots, \varphi(n-1)$ is the number of multisets in $\mathcal{N}[K-k](X)$. Thus:

$$\begin{aligned}
\left|\mathcal{N}[K](X\cup\{y\})\right| &= \sum_{0\leq k\leq K} \left|\mathcal{N}[K-k](X)\right| \\
&\overset{(IH)}{=} \sum_{0\leq k\leq K} \left(\!\binom{n}{K-k}\!\right) \\
&= \left(\!\binom{n+1}{K}\!\right), \qquad \text{by Lemma 1.7.2.} \qquad \text{❑}
\end{aligned}$$

There is also a visual proof of this result, described in terms of stars and bars, see *e.g.* [47, II, proof of (5.2)], where multiplicities of multisets are described in terms of 'occupancy numbers'.

An associated question is: given a fixed element a in an n-element set X, what is the sum of all multiplicities $\varphi(a)$, for multisets φ over X with size K?

Lemma 1.7.5. *For an element $a \in X$, where X has $n \geq 1$ elements,*

$$\sum_{\varphi \in \mathcal{N}[K](X)} \varphi(a) = \frac{K}{n} \cdot \left(\!\binom{n}{K}\!\right).$$

Proof. When we sum over a we get by Proposition 1.7.4:

$$\sum_{a \in X} \sum_{\varphi \in \mathcal{N}[K](X)} \varphi(a) = \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{a \in X} \varphi(a) = \sum_{\varphi \in \mathcal{N}[K](X)} K = K \cdot \left(\!\binom{n}{K}\!\right).$$

Since $a \in X$ is arbitrary, the outcome should be the same for a different $b \in X$. Hence we have to divide by n, giving the equation in the lemma. ❑

We look at one more counting problem, namely the number of multisets below a given multiset. Recall from Definition 1.5.1 (1) the pointwise order $\leq$ on multisets.

Definition 1.7.6. 1 For multisets φ, ψ on the same set X we define the *strict order* $<$ as:

$$\varphi < \psi \iff \varphi \leq \psi \text{ and } \varphi \neq \psi.$$

We say that φ is *fully below* ψ, written as $\varphi \prec \psi$ when for each element the number of occurrences in φ is strictly below (less) than the number of occurrences in ψ. Thus:

$$\varphi \prec \psi \iff \forall x \in X.\, \varphi(x) < \psi(x).$$

2 For $\varphi \in \mathcal{N}(X)$ we use downset notation in the following way:

$$\downarrow \varphi = \{\psi \in \mathcal{N}(X) \mid \psi \leq \varphi\} \qquad \text{and} \qquad \ddagger \varphi = \{\psi \in \mathcal{N}(X) \mid \psi < \varphi\}.$$

Notice that $<$ is not the same as $\prec$, as shown by this example:

$$\begin{aligned} 2|a\rangle + 1|b\rangle + 2|c\rangle &< 2|a\rangle + 2|b\rangle + 2|c\rangle \\ 2|a\rangle + 1|b\rangle + 2|c\rangle &\not\prec 2|a\rangle + 2|b\rangle + 2|c\rangle. \end{aligned}$$

Proposition 1.7.7. *The number of multisets strictly below a natural multiset φ can be described as:*

$$\left| \ddagger \varphi \right| = \sum_{\emptyset \neq U \subseteq supp(\varphi)} \prod_{x \in U} \varphi(x). \tag{1.38}$$

The proof below lends itself to an easy implementation, see Exercise 1.7.12 below. It uses a (chosen) order on the elements of the support of the multiset. The above formulation shows that the outcome is independent of such an order.

Proof. We use induction on the size $|supp(\varphi)|$ of the support of φ. If this size is 1, then φ is of the form $m|x_1\rangle$, for some number m. Hence there are m multisets strictly below φ, namely $\mathbf{0}$, $1|x_1\rangle$, ..., $(m-1)|x_1\rangle$. Since $supp(\varphi) = \{x_1\}$, the only non-empty subset $U \subseteq supp(\varphi)$ is the singleton $U = \{x_1\}$ and $\prod_{x\in U} \varphi(x) = \varphi(x_1) = m$.

Next assume that $supp(\varphi) = X \cup \{y\}$ with $y \notin X$ is of size $n+1$. We write $m = \varphi(y)$ and $\varphi' = \varphi - m|y\rangle$ so that $supp(\varphi') = X$. The number $M = |\, \sharp\varphi' \,|$ is then given by the formula in (1.38), with φ' instead of φ. We count the multisets strictly below φ as follows.

- For each $0 \leq i \leq m$ each $\psi < \varphi'$ gives $\psi + i|y\rangle < \varphi$; this gives $N \cdot (m+1)$ multisets below φ;
- But for each $0 \leq i < m$ also $\varphi' + i|y\rangle$, which gives another m multisets below φ.

We thus get $(m+1) \cdot N + m$ multisets φ, so that:

$$\Big|\, \sharp\varphi \,\Big| \;=\; (\varphi(y)+1) \cdot \Big|\, \sharp\varphi' \,\Big| + \varphi(y). \tag{$*$}$$

Now let's look at the subset formulation in (1.38). Each non-empty subset $U \subseteq X \cup \{y\}$ is either a subset of X, or of the form $U = V \cup \{y\}$ with $V \subseteq X$ non-empty, or contains only y. Hence:

$$\begin{aligned}
\sum_{\emptyset\neq U\subseteq supp(\varphi)} \prod_{x\in U} \varphi(x) \;&=\; \left(\sum_{\emptyset\neq U\subseteq X} \prod_{x\in U} \varphi(x)\right) + \left(\sum_{\emptyset\neq U\subseteq X} \varphi(y) \cdot \prod_{x\in U} \varphi(x)\right) + \varphi(y) \\
&=\; \left(\sum_{\emptyset\neq U\subseteq X} \prod_{x\in U} \varphi'(x)\right) + \varphi(y) \cdot \left(\sum_{\emptyset\neq U\subseteq X} \prod_{x\in U} \varphi'(x)\right) + \varphi(y) \\
&\overset{\text{(IH)}}{=}\; \Big|\, \sharp\varphi' \,\Big| + \varphi(y) \cdot \Big|\, \sharp\varphi' \,\Big| + \varphi(y) \\
&\overset{(*)}{=}\; \Big|\, \sharp\varphi \,\Big|.
\end{aligned}$$

❑

Exercises

1.7.1 Check that $\mathcal{N}[K](1)$ has one element by Proposition 1.7.4. Describe this element. How many elements are there in $\mathcal{N}[K](2)$? Describe them all.

1.7.2 Let φ, ψ be natural multiset on the same finite set X.

1 Show that:

$$\varphi < \psi \iff \varphi + \mathbf{1} \leq \psi \iff \varphi \leq \psi - \mathbf{1},$$

where $\mathbf{1} = \sum_{x\in X} 1|x\rangle$ is the multiset of singletons on X.

2 Now let $\psi \geq \mathbf{1}$. Show that one has:

$$\left(\!\binom{\psi}{\varphi}\!\right) = \binom{\psi + \varphi - \mathbf{1}}{\varphi}.$$

3 Conclude, analogously to Lemma 1.6.4 (1), that:

$$\frac{(\varphi) \cdot (\psi - \mathbf{1})}{(\psi + \varphi - \mathbf{1})} = \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)}.$$

1.7.3 Let $N \geq 0$ and $M \geq m \geq 1$ and be given. Use the multichoose Vandermonde Equation (1.37) to prove:

$$\sum_{0 \leq i \leq N} \left(\!\binom{m}{i}\!\right) \cdot \binom{M - m + N - i}{N - i} = \binom{N + M}{N}.$$

1.7.4 Let $n \geq 1$ and $m \geq 0$.

1 Show that:

$$\left(\!\binom{n+1}{m+1}\!\right) = \left(\!\binom{n+1}{m}\!\right) + \left(\!\binom{n}{m+1}\!\right).$$

2 Generalise this to:

$$\left(\!\binom{n+k}{m+k}\!\right) = \sum_{0 \leq i \leq k} \binom{k}{i} \cdot \left(\!\binom{n+i}{m+k-i}\!\right).$$

1.7.5 Prove the following properties.

1 $\left(\!\binom{n-k}{m-(k+1)}\!\right) = \left(\!\binom{m-k}{n-(k+1)}\!\right)$

2 $\left(\!\binom{n}{m}\!\right) + \left(\!\binom{m}{n}\!\right) = \binom{n+m}{n}$

3 $m \cdot \left(\!\binom{n}{m}\!\right) = n \cdot \left(\!\binom{n+1}{m-1}\!\right)$.

4 $n \cdot \left(\!\binom{n+1}{m}\!\right) = (n+m) \cdot \left(\!\binom{n}{m}\!\right) = (m+1) \cdot \left(\!\binom{n}{m+1}\!\right)$.

1.7.6 1 Show that for numbers $m \leq n - 1$,

$$n \cdot \binom{n-1}{m} = (m+1) \cdot \binom{n}{m+1} = (n-m) \cdot \binom{n}{m}.$$

2 Show similarly that for natural multisets φ, ψ with $x \in \mathit{supp}(\psi)$ and $\varphi \leq \psi - 1|x\rangle$,

$$\psi(x) \cdot \binom{\psi - 1|x\rangle}{\varphi} = (\varphi(x)+1) \cdot \binom{\psi}{\varphi + 1|x\rangle} = (\psi(x) - \varphi(x)) \cdot \binom{\psi}{\varphi}.$$

3 For $x \in \mathit{supp}(\varphi) \subseteq \mathit{supp}(\psi)$,

$$\varphi(x) \cdot \left(\!\binom{\psi}{\varphi}\!\right) = \psi(x) \cdot \left(\!\binom{\psi + 1|x\rangle}{\varphi - 1|x\rangle}\!\right).$$

4 For $supp(\varphi) \subseteq supp(\psi)$ and $x \in supp(\psi)$,

$$\psi(x) \cdot \left(\!\!\binom{\psi+1|\,x\rangle}{\varphi}\!\!\right) = (\varphi(x)+\psi(x)) \cdot \left(\!\!\binom{\psi}{\varphi}\!\!\right) = (\varphi(x)+1) \cdot \left(\!\!\binom{\psi}{\varphi+1|\,x\rangle}\!\!\right).$$

1.7.7 Let $n \geq 1$ and $m \geq 1$.

1 Show that:

$$\sum_{j<m} \left(\!\!\binom{n}{j}\!\!\right) = \left(\!\!\binom{m}{n}\!\!\right).$$

2 Prove next:

$$\sum_{i<n} \left(\!\!\binom{m}{i}\!\!\right) + \sum_{j<m} \left(\!\!\binom{n}{j}\!\!\right) = \binom{n+m}{n}.$$

1.7.8 1 Extend Exercise 1.3.5 (2) to: for $k \geq 1$,

$$\sum_{0\leq j\leq m} \binom{k+n+j}{n} = \left(\!\!\binom{k+m+1}{n+1}\!\!\right) - \left(\!\!\binom{k}{n+1}\!\!\right).$$

Hint: One can use induction on m.

2 Show that one also has:

$$\sum_{0\leq i\leq n} \left(\!\!\binom{k}{i}\!\!\right) \cdot \binom{m+1+n-i}{m} = \left(\!\!\binom{k+m+1}{n+1}\!\!\right) - \left(\!\!\binom{k}{n+1}\!\!\right).$$

Hint: Use the multichoose version (1.36) of Vandermonde's formula.

1.7.9 Show that for $0 < k \leq m+1$,

$$\sum_{i<k} \left(\!\!\binom{k-i}{i}\!\!\right) \cdot \binom{n+m-k+1}{m-i} = \binom{n+m}{n}.$$

1.7.10 Check that Theorem 1.5.2 (1) can be reformulated as: for a real number $x \in (0,1)$ and $K \geq 1$,

$$\sum_{n\geq 0} \left(\!\!\binom{K}{n}\!\!\right) \cdot x^n = \frac{1}{(1-x)^K}$$

1.7.11 Let $s \in [0,1]$ and $n, m \geq 1$.

1 Prove first the following auxiliary result.

$$\sum_{j<m} \left(\!\!\binom{n+1}{j}\!\!\right) \cdot s^j = \frac{1}{(1-s)} \cdot \left(\sum_{j<m} \left(\!\!\binom{n}{j}\!\!\right) \cdot s^j - \left(\!\!\binom{n+1}{m-1}\!\!\right) \cdot s^m \right).$$

2 Take $r = 1 - s$ so that $r + s = 1$ and prove:

$$r^n \cdot \sum_{j<m} \left(\!\binom{n}{j}\!\right) \cdot s^j \;+\; s^m \cdot \sum_{i<n} \left(\!\binom{m}{i}\!\right) \cdot r^i \;=\; 1.$$

3 Show also that:

$$\sum_{i\geq 0} \left(\!\binom{n}{m+i}\!\right) \cdot s^i \;+\; \left(\!\binom{m}{n+i}\!\right) \cdot r^i \;=\; \frac{1}{r^n \cdot s^m}.$$

1.7.12 Let φ be a natural multiset with support $\{x_1, \ldots, x_n\}$. We assume that this support is ordered as a list $[x_1, \ldots, x_n]$. Use the proof of Proposition 1.7.7 to show that the number $|\, \ddagger \varphi|$ of multisets strictly below φ can be computed via the following algorithm.

```
result = 0
for x ∈ [x_1, ..., x_n] :
    result := (φ(x) + 1) ∗ result + φ(x)
return result
```

1.8 Channels

The previous sections covered the collection types of lists, subsets, and multisets, with much emphasis on the similarities between them. In this section we will exploit these similarities in order to introduce the concept of channel, in a uniform approach, for all of these collection types at the same time. This will show that these data types are not only used for certain types of collections, but also for certain types of computation. Much of the rest of this book builds on the concept of a channel, especially for probabilistic distributions, which are introduced in the next chapter. The same general approach to channels that will be described in this section will work for distributions.

Let T be one of the collection functors $\mathcal{L}$, $\mathcal{P}$, or $\mathcal{M}$. A *state* of type Y is an element $\omega \in T(Y)$; it collects a number of elements of Y in a certain way. In this section we abstract away from the particular type of collection. A *channel*, or sometimes more explicitly *T-channel*, is a collection of states, parameterised by a set. Thus, a channel is a function of the form $c\colon X \to T(Y)$. Such a channel turns an element $x \in X$ into a certain collection $c(x)$ of elements of Y. Just like an ordinary function $f\colon X \to Y$ can be seen as a computation, we see a T-channel as a computation *of type* T. For instance, a channel $X \to \mathcal{P}(Y)$ is a non-deterministic computation and a channel $X \to \mathcal{M}(Y)$ is a resource-sensitive computation.

When it is clear from the context what T is, we often write a channel using functional notation, as $c\colon X \nrightarrow Y$, with a circle on the shaft of the arrrow. Notice that a channel $1 \nrightarrow X$ from the singleton set $1 = \{0\}$ can be identified with a state on X.

Definition 1.8.1. Let $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$, each of which is functorial, with its own flatten operation, as described in previous sections.

1 For a state $\omega \in T(X)$ and a channel $c\colon X \to T(Y)$ we can form a new state $c \gg\!= \omega$ of type Y. It is defined as:

$$c \gg\!= \omega \coloneqq (\mathit{flat} \circ T(c))(\omega) \qquad \text{where} \qquad T(X) \xrightarrow{T(c)} T(T(Y)) \xrightarrow{\mathit{flat}} T(Y).$$

This operation $c \gg\!= \omega$ is called *state tranformation*, sometimes with additional *along the channel* c. In functional programming it is called *bind*[1].

2 Let $c\colon X \nrightarrow Y$ and $d\colon Y \nrightarrow Z$ be two channels. Then we can compose them and get a new channel $d \odot c\colon X \nrightarrow Z$ via:

$$(d \odot c)(x) \coloneqq d \gg\!= c(x) \qquad \text{so that} \qquad d \odot c = \mathit{flat} \circ T(d) \circ c.$$

We first look at some examples of state transformation.

Example 1.8.2. Take $X = \{a, b, c\}$ and $Y = \{u, v\}$.

1 For $T = \mathcal{L}$ an example of a state $\omega \in \mathcal{L}(X)$ is $\omega = [c, b, b, a]$. An $\mathcal{L}$-channel $f\colon X \to \mathcal{L}(Y)$ can for instance be given by:

$$f(a) = [u, v] \qquad f(b) = [u, u] \qquad f(c) = [v, u, v].$$

State transformation $f \gg\!= \omega$ amounts to 'map list' with f and then flattening. It turns a list of lists into a list, as in:

$$\begin{aligned} f \gg\!= \omega \;=\; \mathit{flat}(\mathcal{L}(f)(\omega)) &= \mathit{flat}([f(c), f(b), f(b), f(a)]) \\ &= \mathit{flat}([[v, u, v], [u, u], [u, u], [u, v]]) \\ &= [v, u, v, u, u, u, u, u, v]. \end{aligned}$$

2 We consider the analogous example for $T = \mathcal{P}$. We thus take as state $\omega = \{a, b, c\}$ and as channel $f\colon X \to \mathcal{P}(Y)$ with:

$$f(a) = \{u, v\} \qquad f(b) = \{u\} \qquad f(c) = \{u, v\}.$$

[1] Our bind notation $c \gg\!= \omega$ differs from the one used in the functional programming language Haskell; there one writes the state first, as in $\omega \gg\!= c$. For us, the channel c acts on the state ω and is thus written before the argument. This is in line with standard notation $f(x)$ in mathematics, for a function f acting on an argument x. Later on, we shall use predicate transformation $c \mathrel{=\!\!<} p$ along a channel, where we also write the channel first, since it acts on the predicate p. Similarly, in categorical logic the corresponding pullback (or substitution) is written as $c^*(p)$ with the channel before the predicate.

Then:

$$\begin{aligned} f \gg\!= \omega \;=\; \mathit{flat}(\mathcal{P}(f)(\omega)) &= \bigcup \{\, f(a), f(b), f(c) \,\} \\ &= \bigcup \{\, \{u, v\}, \{u\}, \{u, v\} \,\} \;=\; \{u, v\}. \end{aligned}$$

3 For multisets, a state in $\mathcal{M}(X)$ could be of the form $\omega = 3|a\rangle + 2|b\rangle + 5|c\rangle$ and a channel $f\colon X \to \mathcal{M}(Y)$ could have:

$$f(a) \;=\; 10|u\rangle + 5|v\rangle \qquad f(b) \;=\; 1|u\rangle \qquad f(c) \;=\; 4|u\rangle + 1|v\rangle.$$

We then get as state transformation:

$$\begin{aligned} f \gg\!= \omega \;&=\; \mathit{flat}(\mathcal{M}(f)(\omega)) \\ &=\; \mathit{flat}(3|f(a)\rangle + 2|f(b)\rangle + 5|f(c)\rangle) \\ &=\; \mathit{flat}(3\big|\, 10|u\rangle + 5|v\rangle\,\big\rangle + 2\big|\, 1|u\rangle\,\big\rangle + 5\big|\, 4|u\rangle + 1|v\rangle\,\big\rangle) \\ &=\; 30|u\rangle + 15|v\rangle + 2|u\rangle + 20|u\rangle + 5|v\rangle \\ &=\; 52|u\rangle + 20|v\rangle. \end{aligned}$$

We shall mostly be using multiset — and probabilistic channels, as special case — and so we explicitly describe state transformation $\gg\!=$ in these cases. So let $c\colon X \to \mathcal{M}(Y)$ be an $\mathcal{M}$-channel. Transformation of a state ω of type X can be described as:

$$(c \gg\!= \omega)(y) \;=\; \sum_{x \in X} c(x)(y) \cdot \omega(x). \tag{1.39}$$

Equivalently, we can describe the transformed state $c \gg\!= \omega$ as a formal sum:

$$c \gg\!= \omega \;=\; \sum_{y \in Y} \left(\sum_{x \in X} c(x)(y) \cdot \omega(x) \right) |y\rangle. \tag{1.40}$$

We now prove some general properties about state transformation and about composition of channels, based on the abstract description in Definition 1.8.1.

Lemma 1.8.3. *1 Channel composition $\circ$ has a unit, namely $\mathit{unit}\colon Y \nrightarrow Y$, so that:*

$$\mathit{unit} \circ c \;=\; c \qquad \text{and} \qquad d \circ \mathit{unit} \;=\; d,$$

for all channels $c\colon X \nrightarrow Y$ and $d\colon Y \nrightarrow Z$. Another way to write the second equation is: $d \gg\!= \mathit{unit}(y) = d(y)$.

2 Channel composition $\circ$ is associative:

$$e \circ (d \circ c) \;=\; (e \circ d) \circ c,$$

for all channels $c\colon X \nrightarrow Y$, $d\colon Y \nrightarrow Z$ and $e\colon Z \nrightarrow W$.

3 *State tranformation via a composite channel is the same as two consecutive transformations:*

$$(d \circledcirc c) \gg\!= \omega = d \gg\!= (c \gg\!= \omega).$$

4 *Each ordinary function $f\colon Y \to Z$ gives rise to a 'trivial' or 'deterministic' channel $\langle f \rangle := \mathsf{unit} \circ f\colon Y \nrightarrow Z$. This construction $\langle - \rangle$ satisfies:*

$$\langle f \rangle \gg\!= \omega = T(f)(\omega),$$

where T is the type of channel involved. Moreover:

$$\langle g \rangle \circledcirc \langle f \rangle = \langle g \circ f \rangle \qquad \langle f \rangle \circledcirc c = T(f) \circ c \qquad d \circledcirc \langle f \rangle = d \circ f,$$

for all functions $g\colon Z \to W$ and channels $c\colon X \nrightarrow Y$ and $d\colon Y \nrightarrow W$.

Proof. We can give generic proofs, without knowing the type $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$ of the channel, by using earlier results like Lemma 1.2.5, 1.3.2, and 1.4.3. Notice that we carefully distinguish channel composition $\circledcirc$ and ordinary function composition $\circ$.

1 Both equations follow from the *flat − unit* law. By Definition 1.8.1 (2):

$$\mathsf{unit} \circledcirc c = \mathsf{flat} \circ T(\mathsf{unit}) \circ c = \mathsf{id} \circ c = c.$$

For the second equation we use naturality of *unit* in:

$$d \circledcirc \mathsf{unit} = \mathsf{flat} \circ T(d) \circ \mathsf{unit} = \mathsf{flat} \circ \mathsf{unit} \circ d = \mathsf{id} \circ d = d.$$

2 The proof of associativity uses naturality and also the commutation of flatten with itself (the '*flat − flat* law'), expressed as $\mathsf{flat} \circ \mathsf{flat} = \mathsf{flat} \circ T(\mathsf{flat})$.

$$\begin{aligned}
e \circledcirc (d \circledcirc c) &= \mathsf{flat} \circ T(e) \circ (d \circledcirc c) \\
&= \mathsf{flat} \circ T(e) \circ \mathsf{flat} \circ T(d) \circ c \\
&= \mathsf{flat} \circ \mathsf{flat} \circ T(T(e)) \circ T(d) \circ c && \text{by naturality of } \mathsf{flat} \\
&= \mathsf{flat} \circ T(\mathsf{flat}) \circ T(T(e)) \circ T(d) \circ c && \text{by the } \mathsf{flat} - \mathsf{flat} \text{ law} \\
&= \mathsf{flat} \circ T\big(\mathsf{flat} \circ T(e) \circ d\big) \circ c && \text{by functoriality of } T \\
&= \mathsf{flat} \circ T\big(e \circledcirc d\big) \circ c \\
&= (e \circledcirc d) \circledcirc c
\end{aligned}$$

3 Along the same lines:

$$\begin{aligned}
(d \circ c) \gg\!= \omega &= (\mathit{flat} \circ T(d \circ c))(\omega) \\
&= (\mathit{flat} \circ T(\mathit{flat} \circ T(d) \circ c))(\omega) \\
&= (\mathit{flat} \circ T(\mathit{flat}) \circ T(T(d)) \circ T(c))(\omega) && \text{by functoriality of } T \\
&= (\mathit{flat} \circ \mathit{flat} \circ T(T(d)) \circ T(c))(\omega) && \text{by the } \mathit{flat} - \mathit{flat} \text{ law} \\
&= (\mathit{flat} \circ T(d) \circ \mathit{flat} \circ T(c))(\omega) && \text{by naturality of } \mathit{flat} \\
&= (\mathit{flat} \circ T(d))\big((\mathit{flat} \circ T(c))(\omega)\big) \\
&= (\mathit{flat} \circ T(d))(c \gg\!= \omega) \\
&= d \gg\!= (c \gg\!= \omega).
\end{aligned}$$

4 All these properties follow from elementary facts that we have seen before:

$$\begin{aligned}
\langle f \rangle \gg\!= \omega &= (\mathit{flat} \circ T(\mathit{unit} \circ f))(\omega) \\
&= (\mathit{flat} \circ T(\mathit{unit}) \circ T(f))(\omega) && \text{by functoriality of } T \\
&= T(f)(\omega) && \text{by a } \mathit{flat} - \mathit{unit} \text{ law} \\
\langle g \rangle \circ \langle f \rangle &= \mathit{flat} \circ T(\mathit{unit} \circ g) \circ \mathit{unit} \circ f \\
&= \mathit{flat} \circ \mathit{unit} \circ (\mathit{unit} \circ g) \circ f && \text{by naturality of } \mathit{unit} \\
&= \mathit{unit} \circ g \circ f && \text{by a } \mathit{flat} - \mathit{unit} \text{ law} \\
&= \langle g \circ f \rangle \\
\langle f \rangle \circ c &= \mathit{flat} \circ T(\mathit{unit} \circ f) \circ c \\
&= \mathit{flat} \circ T(\mathit{unit}) \circ T(f) \circ c && \text{by functoriality of } T \\
&= T(f) \circ c \\
d \circ \langle f \rangle &= \mathit{flat} \circ T(d) \circ \mathit{unit} \circ f \\
&= \mathit{flat} \circ \mathit{unit} \circ d \circ f && \text{by naturality of } \mathit{unit} \\
&= d \circ f && \text{by a } \mathit{flat} - \mathit{unit} \text{ law.} \quad \square
\end{aligned}$$

In the sequel we often omit writing the brackets $\langle - \rangle$ that turn an ordinary function $f\colon X \to Y$ into a channel $\langle f \rangle$. For instance, in a state transformation $f \gg\!= \omega$, it is clear that we use f as a channel, so that the expression should be read as $\langle f \rangle \gg\!= \omega$.

Exercises

1.8.1 For a function $f\colon X \to Y$ define an *inverse image* $\mathcal{P}$-channel $f^{-1}\colon Y \nrightarrow X$ by:

$$f^{-1}(y) \coloneqq \{x \in X \mid f(x) = y\}.$$

Prove that:

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1} \qquad \text{and} \qquad \mathit{id}^{-1} = \mathit{unit}.$$

1.8.2 Notice that a state of type X can be identified with a channel $1 \to T(Y)$ with singleton set 1 as domain. Check that under this identification, state transformation $c \gg\!= \omega$ corresponds to channel composition $c \circ \omega$.

1.8.3 Let $f\colon X \nrightarrow Y$ be a channel.

1 Prove that if f is a $\mathcal{P}_{fin}$-channel, then the state transformation function $f \gg\!= (-)\colon \mathcal{P}_{fin}(X) \to \mathcal{P}_{fin}(Y)$ can also be defined via freeness, namely as the unique function $\overline{f}$ in Proposition 1.3.4.

2 Similarly, show that $f \gg\!= (-) = \overline{f}$ when f is an $\mathcal{M}$-channel, as in Exercise 1.4.9.

1.8.4 1 Describe how (non-deterministic) powerset channels can be reversed, via a bijective correspondence between functions:

$$\frac{X \longrightarrow \mathcal{P}(Y)}{Y \longrightarrow \mathcal{P}(X)}$$

(A description of this situation in terms of 'daggers' will appear in Example 7.8.1.)

2 Show that for finite sets X, Y there is a similar correspondence for multiset channels.

1.9 The role of category theory

The previous sections have highlighted several structural properties of, and similarities between, the collection types list, powerset, multiset — and later on also distribution. By now readers may ask: what is the underlying structure? Surely someone must have axiomatised what makes all of this work!

Indeed, this is called *category theory*. It provides a foundational language for mathematics, which was first formulated in the 1950s by Saunders Mac Lane and Samuel Eilenberg (see the first overview book [118]). Category theory focuses on the structural aspects of mathematics and shows that many mathematical constructions have the same underlying structure. It brings forward many similarities between different areas (see *e.g.* [119]). Category theory has become very useful in (theoretical) computer science too, since it involves a clear distinction between specification and implementation, see books like [7, 10, 113, 138]. We refer to those sources for more information.

The role of category theory in capturing the mathematical essentials and estabilishing connections also applies to probability theory. William Lawvere, another founding father of the area, first worked in this direction. Lawvere himself published little on this approach to probability theory, but his ideas can be

found in *e.g.* the early notes [111]. This line of work was picked up, extended, and published by his PhD student Michèle Giry. Her name continues in the 'Giry monad' $\mathcal{G}$ of continuous probability distributions, see Section **??**. The precise source of the distribution monad $\mathcal{D}$ for discrete probability theory, that will be introduced in Section 2.1 in the next chapter, is less clear, but it can be regarded as the discrete version of $\mathcal{G}$. Probabilistic automata have been studied in categorical terms as coalgebras, see Chapter **??**, and *e.g.* [152] and [70] for general background information on coalgebra. There is a recent surge in interest in more foundational, semantically oriented studies in probability theory, through the rise of probabilistic programming languages [59, 153], probabilistic Bayesian reasoning [28, 89], and category theory **??**. Probabilistic methods have received wider attention, for instance, via the current interest in data analytics (see the essay [4]), in quantum probability [128, 25], and in cognition theory [64, 151].

Readers who know category theory will have recognised its implicit use in earlier sections. For readers who are not familiar (yet) with category theory, some basic concepts will be explained informally in this section. This is in no way a serious introduction to the area. The remainder of this book will continue to make implicit use of category theory, but will make this usage increasingly explicit. Hence it is useful to know the basic concepts of category, functor, natural transformation, and monad. Category theory is sometimes seen as a difficult area to get into. But our experience is that it is easiest to learn category theory by recognising its concepts in constructions that you already know. That is why this chapter started with concrete descriptions of various collections and their use in channels. For more solid expositions of category theory we refer to the sources listed above.

1.9.1 Categories

A category is a mathematical structure given by a collection of 'objects' with 'morphisms' between them. The requirements are that these morphisms are closed under (associative) composition and that there is an identity morphism on each object. Morphisms are also called 'maps', and are written as $f \colon X \to Y$, where X, Y are objects and f is a homomorphism from X to Y. It is tempting to think of morphisms in a category as actual functions, but there are plenty of examples where this is not the case.

A category is like an abstract context of discourse, giving a setting in which one is working, with properties of that setting depending on the category at hand. We shall give a number of examples.

1 There is the category **Sets**, whose objects are sets and whose morphisms are ordinary functions between them. This is a standard example.
2 One can also restrict to finite sets as objects, in the category **FinSets**, with functions between them. This category is more restrictive, since for instance it contains objects $n = \{0, 1, \ldots, n-1\}$ for each $n \in \mathbb{N}$, but not $\mathbb{N}$ itself. Also, in **Sets** one can take arbitrary products $\prod_{i \in I} X_i$ of objects X_i, over arbitrary index sets I, whereas in **FinSets** only finite products exist. Hence **FinSets** is a more restrictive world.
3 Monoids and monoid maps have been mentioned in Definition 1.2.1. They can be organised in a category **Mon**, whose objects are monoids, and whose homomorphisms are monoid maps. We now have to check that monoid maps are closed under composition and that identity functions are monoid maps; this is easy. Many mathematical structures can be organised into categories in this way, where the morphisms preserve the relevant structure. For instance, one can form a category **PoSets**, with partially ordered sets (posets) as objects, and monotone functions between them as morphisms (also closed under composition, with identity).
4 For $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$ we can form the category $\mathbf{Chan}(T)$. Its objects are arbitrary sets X, but its morphisms X to Y are T-channels, $X \to T(Y)$, written as $X \nrightarrow Y$. We have already seen that channels are closed under composition $\circledcirc$ and have *unit* as identity, see Lemma 1.8.3. We can now say that $\mathbf{Chan}(T)$ is a category.

   These categories of channels form good examples of the idea that a category forms a universe of discourse. For instance, in $\mathbf{Chan}(\mathcal{P})$ we are in the world of non-deterministic computation, whereas $\mathbf{Chan}(\mathcal{M})$ is the world of computation in which resources are counted.

We will encounter several more examples of categories later on in the book. Occasionally, the following construction will be used. Given a category $\mathbb{C}$, a new 'opposite' category $\mathbb{C}^{\mathrm{op}}$ is formed. It has the same objects as $\mathbb{C}$, but its morphisms are reversed. Thus $f \colon Y \to X$ in $\mathbb{C}^{\mathrm{op}}$ means $f \colon X \to Y$ in $\mathbb{C}$.

Also, given two categories $\mathbb{C}$, $\mathbb{D}$ one can form a product category $\mathbb{C} \times \mathbb{D}$. Its objects are pairs (X, A) with X an object in $\mathbb{C}$ and A an object in $\mathbb{D}$. Similarly, an arrow $(X, A) \to (Y, B)$ in $\mathbb{C} \times \mathbb{D}$ is given by a pair (f, g) of arrows $f \colon X \to Y$ in $\mathbb{C}$ and $g \colon A \to B$ in $\mathbb{D}$.

1.9.2 Functors

Category theorists like abstraction, hence the question: if categories are so important, then why not organise them as objects themselves in a superlarge cat-

egory **Cat**, with morphisms between them preserving the relevant structure? The latter morphisms between categories are called 'functors'. More precisely, given categories $\mathbb{C}$ and $\mathbb{D}$, a functor $F\colon \mathbb{C} \to \mathbb{D}$ between them consists of two mappings, both written F, sending an object X in $\mathbb{C}$ to an object $F(X)$ in $\mathbb{D}$, and a morphism $f\colon X \to Y$ in $\mathbb{C}$ to a morphism $F(f)\colon F(X) \to F(Y)$ in $\mathbb{D}$. This mapping F should preserve composition and identities, as in: $F(g \circ f) = F(g) \circ F(f)$ and $F(id_X) = id_{F(X)}$.

Earlier we have already called some operations 'functorial' for the fact that they preserve composition and identities. We can now be a bit more precise.

1 Each $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{P}_{fin}, \mathcal{M}, \mathcal{M}_*, \mathcal{N}, \mathcal{N}_*, \mathcal{N}[K]\}$ is a functor $T\colon$ **Sets** $\to$ **Sets**. This has been described in the beginning of each of the sections 1.2 – 2.1.
2 Taking lists is also a functor $\mathcal{L}\colon$ **Sets** $\to$ **Mon**. This is in essence the content of Lemma 1.2.2. One can also view $\mathcal{P}, \mathcal{P}_{fin}$ and $\mathcal{M}$ as functors **Sets** $\to$ **Mon**, see Lemmas 1.3.1 and 1.4.2. Moreover, one can describe $\mathcal{P}, \mathcal{P}_{fin}$ as a functor **Sets** $\to$ **PoSets**, by considering each set of subsets $\mathcal{P}(X)$ and $\mathcal{P}_{fin}(X)$ with its subset relation $\subseteq$ as partial order. In order to verify this claim one has to check that $\mathcal{P}(f)\colon \mathcal{P}(X) \to \mathcal{P}(Y)$ is a morphism of posets, that is, forms a monotone function. But that is easy.
3 There is also a functor $J\colon$ **Sets** $\to$ **Chan**(T), for each T. It is the identity on sets/objects: $J(X) \coloneqq X$. But it sends a functon $f\colon X \to Y$ to the channel $J(f) \coloneqq \langle f \rangle = unit \circ f\colon X \nrightarrow Y$. We have seen, in Lemma 1.8.3 (4), that $J(g \circ f) = J(g) \circ J(f)$ and that $J(id) = id$, where the latter identity id is $unit$ in the category **Chan**(T). This functor J shows how to embed the world of ordinary computations (functions) into the world of compuations of type T (channels).
4 Taking the product of two sets can be described as a functor $\times\colon$ **Sets**$\times$**Sets** $\to$ **Sets**. Its action on morphisms was already described at the end of Subsection 1.1.1, see also Exercise 1.1.3.
5 If we have functors $F_i\colon \mathbb{C}_i \to \mathbb{D}_i$, for $i = 1, 2$, then we also have a product functor $F_1 \times F_2\colon \mathbb{C}_1 \times \mathbb{C}_2 \to \mathbb{D}_1 \times \mathbb{D}_2$ between product categories, simply by $(F_1 \times F_2)(X_1, X_2) = (F_1(X_1), F_2(X_2))$, and similarly for morphisms.

1.9.3 Natural transformations

Let us move one further step up the abstraction ladder and look at morphisms between functors. These are called *natural transformations*. We have already seen examples of those as well. Given two functors $F, G\colon \mathbb{C} \to \mathbb{D}$, a natural transformation α from F to G is a collection of maps $\alpha_X\colon F(X) \to G(X)$ in $\mathbb{D}$, indexed by objects X in $\mathbb{C}$. Naturality means that α works in the same way on

all objects and is expressed as follows: for each morphism $f\colon X \to Y$ in $\mathbb{C}$, the rectangle

$$\begin{array}{ccc} F(X) & \xrightarrow{\alpha_X} & G(X) \\ {\scriptstyle F(f)}\downarrow & & \downarrow{\scriptstyle G(f)} \\ F(Y) & \xrightarrow[\alpha_Y]{} & G(Y) \end{array}$$

in $\mathbb{D}$ commutes.

Such a natural transformation is often denoted by a double arrow $\alpha\colon F \Rightarrow G$. We briefly review some of the examples of natural transformations that we have seen.

1 The various support maps can now be described as natural transformations $\mathit{supp}\colon \mathcal{L} \Rightarrow \mathcal{P}_{\mathit{fin}}$, $\mathit{supp}\colon \mathcal{M} \Rightarrow \mathcal{P}_{\mathit{fin}}$ and $\mathit{supp}\colon \mathcal{D} \Rightarrow \mathcal{P}_{\mathit{fin}}$, see the overview diagram 1.30.
2 For each $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$ we have described maps $\mathit{unit}\colon X \to T(X)$ and $\mathit{flat}\colon T(T(X)) \to T(X)$ and have seen naturality results about them. We can now state more precisely that they are natural transformations $\mathit{unit}\colon \mathit{id} \Rightarrow T$ and $\mathit{flat}\colon (T \circ T) \Rightarrow T$. Here we have used id as the identity functor $\mathbf{Sets} \to \mathbf{Sets}$, and $T \circ T$ as the composite of T with itself, also as a functor $\mathbf{Sets} \to \mathbf{Sets}$.

1.9.4 Monads

A monad on a category $\mathbb{C}$ is a functor $T\colon \mathbb{C} \to \mathbb{C}$ that comes with two natural transformations $\mathit{unit}\colon \mathit{id} \Rightarrow T$ and $\mathit{flat}\colon (T \circ T) \Rightarrow T$ satisfying:

$$\begin{array}{c} \mathit{flat} \circ \mathit{unit} \;=\; \mathit{id} \;=\; \mathit{flat} \circ T(\mathit{unit}) \\ \mathit{flat} \circ \mathit{flat} \;=\; \mathit{flat} \circ T(\mathit{flat}). \end{array} \tag{1.41}$$

All the collection functors $\mathcal{L}, \mathcal{P}, \mathcal{P}_{\mathit{fin}}, \mathcal{M}, \mathcal{M}_*, \mathcal{N}, \mathcal{N}_*, \mathcal{D}, \mathcal{D}_\infty$ that we have seen so far are monads, see *e.g.*, Lemma 1.2.5, 1.3.2, or 1.4.3. For each monad T we can form a category $\mathbf{Chan}(T)$ of T-channels, that capture computations of type T, see Subsection 1.9.1. In category theory this is called the Kleisli category of T. Composition in this category $\mathbf{Chan}(T)$ is called Kleisli composition. In this book it is written as $\circledcirc$, where the context should make clear what the monad T at hand is.

Monads have become popular in functional programming [126] as mechanisms for including special effects (*e.g.*, for input-output, writing, side-effects, continuations) into a functional programming language[2]. The structure of prob-

[2] See the online overview https://wiki.haskell.org/Monad_tutorials_timeline

abilistic computation is also given by monads, namely by the discrete distribution monads $\mathcal{D}, \mathcal{D}_\infty$ and by the continuous distribution monad $\mathcal{G}$.

We thus associate the (Kleisli) category $\mathbf{Chan}(T)$ of channels with a monad T. A second category is associated with a monad T, namely the category $\mathcal{EM}(T)$ of "Eilenberg-Moore" algebras. The objects of $\mathcal{EM}(T)$ are algebras $\alpha\colon T(X) \to X$, satisfying $\alpha \circ \mathit{unit} = \mathit{id}$ and $\alpha \circ \mathit{flat} = \alpha \circ T(\alpha)$. We have seen algebras for the monads $\mathcal{L}, \mathcal{P}_{\mathit{fin}}$, and $\mathcal{M}$ in Propositions 1.2.6, 1.3.5, and 1.4.5. They capture monoids, commutative idempotent monoids, and commutative monoids respectively. A morphism in $\mathcal{EM}(T)$ is a morphism of algebras, given by a commuting rectangle, as described in these propositions. In general, algebras of a monad capture algebraic structure in a uniform manner.

Here is an easy result that describes so-called *writer* monads.

Lemma 1.9.1. *Let $M = (M, +, 0)$ be an arbitrary monoid. The mapping $X \mapsto M \times X$ forms a monad on the category* **Sets**.

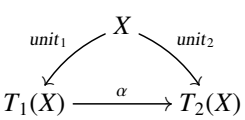
Proof. Let us write $T(X) = M \times X$. For a function $f\colon X \to Y$ we define $T(f)\colon M \times X \to M \times Y$ by $T(f)(m, x) = (m, f(x))$. There is a unit map $\mathit{unit}\colon X \to M \times X$, namely $\mathit{unit}(x) = (0, x)$ and a flattening map $\mathit{flat}\colon M \times (M \times X) \to M \times X$ by $\mu(m, m', x) = (m + m', x)$. We skip naturality and concentrate on the monad equations (1.41). First, for $(m, x) \in T(X) = M \times X$,

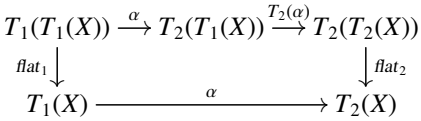
$$\begin{aligned}(\mathit{flat} \circ \mathit{unit})(m, x) &= \mathit{flat}(0, m, x) = (0 + m, x) = (m, x)\\ (\mathit{flat} \circ T(\mathit{unit}))(m, x) &= \mathit{flat}\big(m, \mathit{unit}(x)\big) = \mathit{flat}(m, 0, x) = (m, x).\end{aligned}$$

Next, the flatten-equation holds by associativity of the monoid addition $+$. This is left to the reader. ❑

We have seen natural transformations as maps between functors. In the special case where the functors involved are monads, these natural transformations can be called maps of monads if they additionally commute with the unit and flatten maps.

Definition 1.9.2. Let $T_1 = (T_1, \mathit{unit}_1, \mathit{flat}_1)$ and $T_2 = (T_2, \mathit{unit}_2, \mathit{flat}_2)$ be two monads (on **Sets**). A *map/homomorphism of monads* from T_1 to T_2 is a natural transformation $\alpha\colon T_1 \Rightarrow T_2$ that commutes with unit and flatten in the sense that the two diagrams





commute, for each set X.

The writer monads from Lemma 1.9.1 give simple examples of maps of monads: if $f\colon M_1 \to M_2$ is a map of monoids, then the maps $\alpha \coloneqq f \times id : M_1 \times X \to M_2 \times X$ form a map of monoids.

For a historical account of monads and their applications we refer to [66].

Exercises

1.9.1 We have seen the functor $J\colon \mathbf{Sets} \to \mathbf{Chan}(T)$. Check that there is also a functor $\mathbf{Chan}(T) \to \mathbf{Sets}$ in the opposite direction, which is $X \mapsto T(X)$ on objects, and $c \mapsto c \gg\!= (-)$ on morphisms. Check explicitly that composition is preserved, and find the earlier result that stated that fact implicitly.

1.9.2 Recall from (1.15) the subset $\mathcal{N}[K](X) \subseteq \mathcal{M}(X)$ of natural multisets with K elements. Prove that $\mathcal{N}[K]$ is a functor $\mathbf{Sets} \to \mathbf{Sets}$.

1.9.3 Show that Exercise 1.8.1 implicitly describes a functor $\mathbf{Sets}^{\mathrm{op}} \to \mathbf{Chan}(\mathcal{P})$, which is the identity on objects.

1.9.4 Show that the *zip* function from Exercise 1.1.7 is natural: for each pair of functions $f\colon X \to U$ and $g\colon Y \to V$ the following diagram commutes.

$$\begin{array}{ccc} X^K \times Y^K & \xrightarrow{zip} & (X \times Y)^K \\ {\scriptstyle f^K \times g^K}\downarrow & & \downarrow{\scriptstyle (f\times g)^K} \\ U^K \times V^K & \xrightarrow{zip} & (U \times V)^K \end{array}$$

1.9.5 Fill in the remaining details in the proof of Lemma 1.9.1: that T is a functor, that *unit* and *flat* are natural transformation, and that the flatten equation holds.

1.9.6 For arbitrary sets X, A, write $X + A$ for the disjoint union (coproduct) of X and A, which may be described explicitly by tagging elements with $1, 2$ in order to distinguish them:

$$X + A \;=\; \{(x, 1) \mid x \in X\} \cup \{(a, 2) \mid a \in A\}.$$

Write $\kappa_1 : X \to X + A$ and $\kappa_2 : A \to X + A$ for the two obvious functions.

1 Keep the set A fixed and show that the mapping $X \mapsto X + A$ can be extended to a functor $\mathbf{Sets} \to \mathbf{Sets}$.

2 Show that it is actually a monad; it is sometimes called the *exception* monad, where the elements of A are seen as exceptions in a computation.

1.9.7 Check that the support and accumulation functions form maps of monads in the situations:

1 $supp \colon \mathcal{M}(X) \Rightarrow \mathcal{P}(X)$;
2 $acc \colon \mathcal{L}(X) \Rightarrow \mathcal{M}(X)$.

1.9.8 Let $T = (T, unit, flat)$ be a monad. By definition, it involves T as a functor $T \colon \mathbf{Sets} \to \mathbf{Sets}$. Show that T can be 'lifted' to a functor $\overline{T} \colon \mathbf{Chan}(T) \to \mathbf{Chan}(T)$. It is defined on objects as $\overline{T}(X) \coloneqq T(X)$ and on a morphism $f \colon X \nrightarrow Y$ as:

$$\overline{T}(f) \coloneqq \Big(T(X) \xrightarrow{T(f)} T(T(Y)) \xrightarrow{flat} T(Y) \xrightarrow{unit} T(T(Y)) \Big).$$

Prove that $\overline{T}$ is a functor, *i.e.* that it preserves (channel) identities and composition.

2

Discrete probability distributions

The previous chapter has introduced products, lists, subsets and multisets as basic collection types and has made some of their basic properties explicit. This serves as preparation for the current first chapter on probability distributions. We shall see that distributions also form a collection type, with much analogous structure. In fact distributions are special multisets, where multiplicities add up to one.

This chapter introduces the basics of probability distributions and of probabilistic channels (as indexed / conditional distributions). These notions will play a central role in the rest of this book. Distributions will be defined as special multisets, so that there is a simple inclusion of the set of distributions in the set of multisets multisets, on a particular space. In the other direction, this chapter describes the 'frequentist learning' construction, which turns a multiset into a distribution, essentially by normalisation. The chapter also describes several constructions on distributions, like the convex sum, parallel product, and addition of distributions (in the special case when the underlying space happens to be a commutative monoid). Parallel products $\otimes$ of distributions and joint distributions (on product spaces) are rather special, for instance, because a joint distribution is typically not equal to the product of its marginals. Indeed, joint distributions may involve correlations between the different (product) components, so that updates in one component have 'crossover' effect in other components. This magical phenonenom will be elaborated in later chapters.

The chapter closes with an example of a Bayesian network. It shows that the conditional probability tables that are associated with nodes in a Bayesian network are instances of probabilistic channels. As a result, one can systematically organise computations in Bayesian networks as suitable (sequential and/or parallel) compositions of channels. This is illustrated via calculations of various predicted probabilities.

2.1 Probability distributions

This section is the first one about probability, in elementary form. It introduces finite discrete probability distributions, which we often simply call distributions or states. In the literature they are also called 'multinomial' or 'categorical' distributions. The notation and definitions that we use for distributions are very much like for multisets, since distributions form a subset of multisets, namely where multiplicities add up to one.

What we call *distribution*, is in fact a finite discrete probability distribution. We will use *state* as synonym for *distribution*. A distribution over a set X is a finite formal convex sum of the form:

$$r_1|x_1\rangle + \cdots + r_n|x_n\rangle \qquad \text{where} \qquad x_i \in X \text{ and } r_i \in [0,1] \text{ with } \sum_i r_i = 1.$$

We can write such an expression as a sum $\sum_i r_i|x_i\rangle$. It is called a *convex* sum since the r_i add up to one. Thus, a distribution is a special 'probabilistic' multiset.

We write $\mathcal{D}(X)$ for the set of distributions on a space / set X, so that there is an inclusion $\mathcal{D}(X) \subseteq \mathcal{M}(X)$ of distributions on X in the set of multisets on X. Via this inclusion, we use the same conventions for distributions, as for multisets; they were described in the three bullet points in the beginning of Section 1.4. This set X is often called the *sample space*, see *e.g.* [144], the *outcome space*, the *underlying space*, or simply the *underlying set*. Each element $x \in X$ gives rise to a distribution $1|x\rangle \in \mathcal{D}(X)$, which is 1 on x and 0 everywhere else. It is called a Dirac distribution, a point mass, or also a point distribution. The mapping $x \mapsto 1|x\rangle$ is the unit function $\mathit{unit} \colon X \to \mathcal{D}(X)$.

For a coin we can use the set $\{H, T\}$ with elements for head and tail as sample space. A fair coin is described on the left below, as a distribution over this set; the distribution on the right gives a coin with a slight bias.

$$\tfrac{1}{2}|H\rangle + \tfrac{1}{2}|T\rangle \qquad\qquad 0.51|H\rangle + 0.49|T\rangle.$$

In general, for a finite set $X = \{x_1, \ldots, x_n\}$ there a *uniform* distribution $\mathit{unif}_X \in \mathcal{D}(X)$ that assigns the same probability to each element. Thus, it is given by $\mathit{unif}_X = \sum_{1 \leq i \leq n} \frac{1}{n}|x_i\rangle$. The above fair coin is a uniform distribution on the two-element set $\{H, T\}$. Similarly, a fair dice can be described as $\mathit{unif}_{\mathit{pips}} = \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$, where $\mathit{pips} = \{1, 2, 3, 4, 5, 6\}$. Figure 2.1 shows bar charts of several distributions. The last one describes the letter frequencies in English for the latin alphabet. One commonly does not distinguish upper en lower cases in such frequencies, so we take the 26-element set $A = \{a, b, c, \ldots, z\}$ of lower cases as sample space. The distribution itself

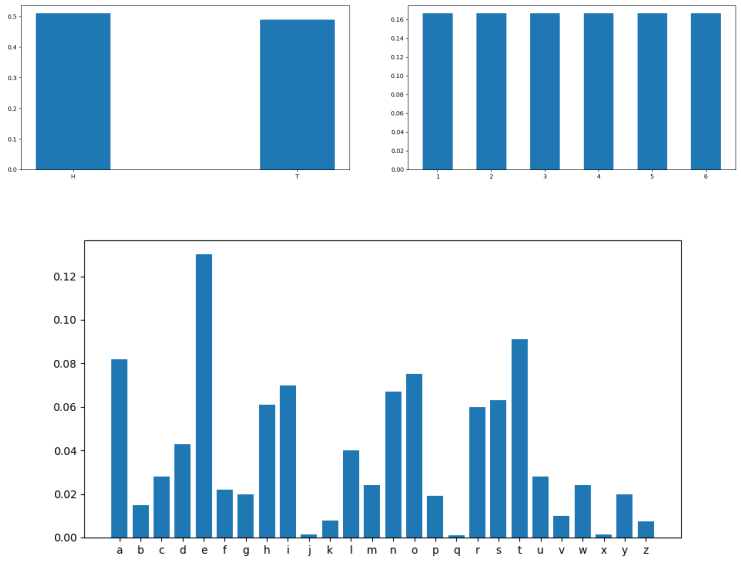


Figure 2.1 Plots of a slightly biased coin distribution $0.51|H\rangle + 0.49|T\rangle$ and a fair (uniform) dice distribution on $\{1, 2, 3, 4, 5, 6\}$ in the top row, together with the distribution of letter frequencies in English at the bottom.

can be described as formal sum:

$$\begin{aligned}
&0.082|a\rangle + 0.015|b\rangle + 0.028|c\rangle + 0.043|d\rangle + 0.13|e\rangle + 0.022|f\rangle \\
&\quad + 0.02|g\rangle + 0.061|h\rangle + 0.07|i\rangle + 0.0015|j\rangle + 0.0077|k\rangle \\
&\quad + 0.04|l\rangle + 0.024|m\rangle + 0.067|n\rangle + 0.075|o\rangle + 0.019|p\rangle \\
&\quad + 0.00095|q\rangle + 0.06|r\rangle + 0.063|s\rangle + 0.091|t\rangle + 0.028|u\rangle \\
&\quad + 0.0098|v\rangle + 0.024|w\rangle + 0.0015|x\rangle + 0.02|y\rangle + 0.0074|z\rangle.
\end{aligned}$$

These frequencies have been copied from Wikipedia. Interestingly, they do not precisely add up to 1, but to 1.01085, probably due to rounding. Thus, strictly speaking, this is not a probability distribution but a multiset.

Below we describe several standard examples of distributions — which will play an important role in the rest of the book. Especially the three 'draw' distributions — multinomial, hypergeometric, Pólya — capture probabilities associated with drawing coloured balls from an urn. The differences between these distributions involve the changes in the urn after the draws and can be characterised as -1, 0, and $+1$. In the hypergeometric case a drawn ball is removed (-1), in the multinomial case a drawn ball is returned so that the urn

remains unaltered (0), and in the Pólya case the drawn ball is returned to the urn together with an extra ball of the same colour (+1).

Example 2.1.1. We shall explicitly describe several familiar distributions using the above formal convex sum notation.

1 The coin that we have seen above can be parametrised via a 'bias' probability $r \in [0, 1]$. The resulting coin is often called *flip* and is defined as:

$$\mathit{flip}(r) \coloneqq r|1\rangle + (1 - r)|0\rangle$$

where 1 may understood as 'head' and 0 as 'tail'. We may thus see *flip* as a function $\mathit{flip}\colon [0, 1] \to \mathcal{D}(\{0, 1\})$ from probabilities to distributions over the sample space $2 = \{0, 1\}$ of Booleans. This $\mathit{flip}(r)$ is often called the *Bernoulli* distribution, with parameter $r \in [0, 1]$.

2 For each number $K \in \mathbb{N}$ and probability $r \in [0, 1]$ there is the familiar *binomial* distribution $\mathit{bn}[K](r) \in \mathcal{D}(\{0, 1, \ldots, K\})$. It captures probabilities for iterated coin flips, and is given by the convex sum:

$$\mathit{bn}[K](r) \coloneqq \sum_{0 \leq k \leq K} \binom{K}{k} \cdot r^k \cdot (1 - r)^{K-k} \,\big|k\big\rangle.$$

The multiplicity probability before $|k\rangle$ in this expression is the chance of getting k heads of out K coin flips, where each flip has bias $r \in [0, 1]$. In this way we obtain a function $\mathit{bn}[K]\colon [0, 1] \to \mathcal{D}(\{0, 1, \ldots, K\})$. These multiplicities are plotted as bar charts in the top row of Figure 2.2, for two binomial distributions, both on the sample space $\{0, 1, \ldots, 10\}$.

There are isomorphisms $[0, 1] \cong \mathcal{D}(2)$, via the above *flip* function, and $\{0, 1, \ldots, K\} \cong \mathcal{N}[K](2)$, via $k \mapsto k|0\rangle + (K - k)|1\rangle$. With these isomorphisms we can describe binomials as probabilistic function $\mathit{bn}[K]\colon \mathcal{D}(2) \to \mathcal{D}(\mathcal{N}[K](2))$. This re-formulation opens the door to a multivariate version of this binomial distribution, called the *multinomial* distribution. It can be described as function:

$$\mathcal{D}(X) \xrightarrow{\quad \mathit{mn}[K] \quad} \mathcal{D}\Big(\mathcal{N}[K](X)\Big). \tag{2.1}$$

The number $K \in \mathbb{N}$ represents the number of objects that is drawn, from a distribution $\omega \in \mathcal{D}(X)$, seen as abstract urn. The distribution $\mathit{mn}[K](\omega)$ assigns a probability to a K-sized draw $\varphi \in \mathcal{N}[K](X)$. There is no bound on K, since the idea behind multinomial distributions is that drawn objects are replaced. More details will appear in Chapter 3; at this stage it suffices to define for $\omega \in \mathcal{D}(X)$,

$$\mathit{mn}[K](\omega) \coloneqq \sum_{\varphi \in \mathcal{N}[K](X)} \binom{\varphi}{} \cdot \omega^{\varphi} \,\big|\varphi\big\rangle, \tag{2.2}$$

where $\binom{\varphi}{}$ is the multinomial coefficient $\frac{K!}{\prod_x \varphi(x)!}$, see Definition 1.5.1 (5) and where $\omega^\varphi := \prod_x \omega(x)^{\varphi(x)}$. In the sequel we shall standardly use the multinomial distribution and view the binomial distribution as a special case.

To see an example, for space $X = \{a, b, c\}$ and urn $\omega = \frac{1}{3}|a\rangle + \frac{1}{2}|b\rangle + \frac{1}{6}|c\rangle$ the draws of size 3 form a distribution over multisets, described below within the outer 'big' kets $\big|-\big\rangle$.

$$\begin{aligned} mn[3](\omega) = \; & \tfrac{1}{27}\big|3|a\rangle\big\rangle + \tfrac{1}{6}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{8}\big|3|b\rangle\big\rangle \\ & + \tfrac{1}{18}\big|2|a\rangle + 1|c\rangle\big\rangle + \tfrac{1}{6}\big|1|a\rangle + 1|b\rangle + 1|c\rangle\big\rangle + \tfrac{1}{8}\big|2|b\rangle + 1|c\rangle\big\rangle \\ & + \tfrac{1}{36}\big|1|a\rangle + 2|c\rangle\big\rangle + \tfrac{1}{24}\big|1|b\rangle + 2|c\rangle\big\rangle + \tfrac{1}{216}\big|3|c\rangle\big\rangle \end{aligned}$$

Via the Multinomial Theorem (1.27) we see that the probabilities in the above expression (2.2) for $mn[K](\omega)$ add up to one:

$$\sum_{\varphi \in \mathcal{N}[K](X)} \binom{\varphi}{} \cdot \omega^\varphi = \sum_{\varphi \in \mathcal{N}[K](X)} \binom{\varphi}{} \cdot \prod\nolimits_x \omega(x)^{\varphi(x)} \overset{(1.26)}{=} \Big(\sum\nolimits_x \omega(x)\Big)^K = 1^K = 1.$$

In addition, we note that the multisets φ in (2.2) can be restricted to those with $supp(\varphi) \subseteq supp(\omega)$. Indeed, if $\omega(x) = 0$, but $\varphi(x) \neq 0$, for some $x \in X$, then $\omega(x)^{\varphi(x)} = 0$, so that the whole product $\prod$ becomes zero, and so that φ does not contribute to the above multinomial distribution.

3 The multinomial distribution captures draws *with replacement*, where drawn balls are returned to the urn. Hence the urn itself can be described as a probability distribution. A useful variation is the *hypergeometric* distribution which captures draws from an urn *without replacement*. We briefly introduce this hypergeometric distribution, in multivariate form, and postpone further analysis to Chapter 3.

When drawn balls are not replaced, the urn in question changes with each draw. In the hypergeometric case the urn is a multiset, say of size $L \in \mathbb{N}$. Draws are then multisets of size K, with $K \leq L$. The hypergeometric function thus takes the form:

$$\mathcal{N}[L](X) \xrightarrow{\quad hg[K] \quad} \mathcal{D}(\mathcal{N}[K](X)). \tag{2.3}$$

This function/channel is defined on an L-sized multiset $\psi \in \mathcal{N}[L](X)$ as:

$$hg[K](\psi) := \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L}{K}} \big|\varphi\big\rangle = \sum_{\varphi \leq_K \psi} \frac{\prod_x \binom{\psi(x)}{\varphi(x)}}{\binom{L}{K}} \big|\varphi\big\rangle.$$

This yields a probability distribution by Lemma 1.6.2. To see an example,

with $\psi = 4|a\rangle + 6|b\rangle + 2|c\rangle$ as urn, the draws of size 3 give a distribution:

$$\begin{aligned} hg[3](\psi) = \ & \tfrac{1}{55}\big|3|a\rangle\big\rangle + \tfrac{9}{55}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{3}{11}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{11}\big|3|b\rangle\big\rangle \\ & + \tfrac{3}{55}\big|2|a\rangle + 1|c\rangle\big\rangle + \tfrac{12}{55}\big|1|a\rangle + 1|b\rangle + 1|c\rangle\big\rangle + \tfrac{3}{22}\big|2|b\rangle + 1|c\rangle\big\rangle \\ & + \tfrac{1}{55}\big|1|a\rangle + 2|c\rangle\big\rangle + \tfrac{3}{110}\big|1|b\rangle + 2|c\rangle\big\rangle. \end{aligned}$$

4 We have seen that in multinomial mode the drawn ball is returned to the urn, whereas in hypergeometric mode the drawn ball is removed. There is a logical third option where the drawn ball is returned, together with one additional ball of the same colour. This leads to what is called the Pólya distribution. We shall describe it as a function of the form:

$$\mathcal{N}_*(X) \xrightarrow{\quad pl[K] \quad} \mathcal{D}(\mathcal{N}[K](X)). \tag{2.4}$$

Its description is very similar to the hypergeometric distribution, with as main difference that it uses multichoose binomomial coefficients instead of ordinary binomial coefficients:

$$pl[K](\psi) = \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{\|\psi\|}{K}\!\right)} \big|\varphi\big\rangle. \tag{2.5}$$

Notices that the draws φ satisfy $supp(\varphi) \subseteq supp(\psi)$. In this situation it is most natural to use urns ψ with full support, that is with $\psi(x) > 0$ for each $x \in X$. In that case the urn contains at least one ball of each colour.

The above Pólya formula in (2.5) yields a proper distribution by Proposition 1.7.3. Here is an example, with the same urn $\psi = 4|a\rangle + 6|b\rangle + 2|c\rangle$ as before, for the hypergeometric distribution.

$$\begin{aligned} pl[3](\psi) = \ & \tfrac{5}{91}\big|3|a\rangle\big\rangle + \tfrac{15}{91}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{3}{13}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{2}{13}\big|3|b\rangle\big\rangle \\ & + \tfrac{5}{91}\big|2|a\rangle + 1|c\rangle\big\rangle + \tfrac{12}{91}\big|1|a\rangle + 1|b\rangle + 1|c\rangle\big\rangle + \tfrac{3}{26}\big|2|b\rangle + 1|c\rangle\big\rangle \\ & + \tfrac{3}{91}\big|1|a\rangle + 2|c\rangle\big\rangle + \tfrac{9}{182}\big|1|b\rangle + 2|c\rangle\big\rangle + \tfrac{1}{91}\big|3|c\rangle\big\rangle \end{aligned}$$

The middle and last rows in Figure 2.2 give bar plots for hypergeometric and Pólya distributions over $2 = \{0, 1\}$. They show the probabilities for numbers $0 \leq k \leq 10$ in a drawn multiset $k|0\rangle + (10 - k)|1\rangle$.

So far we have described distributions as formal convex sums. But they can be described, equivalently, in functional form. This is done in the definition below, which is much like Definition 1.4.1 for multisets. Also for distributions we shall freely switch between the above ket-formulation and the function-formulation.

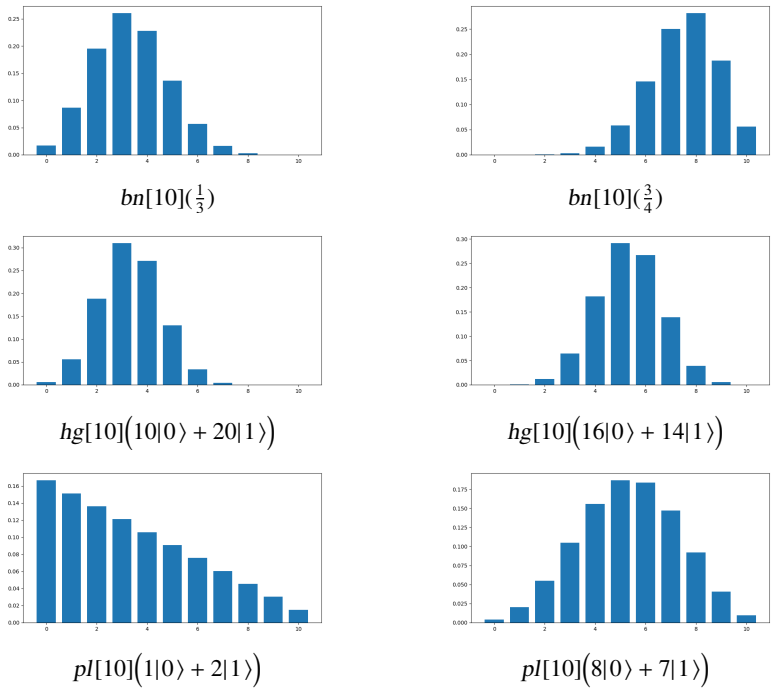


Figure 2.2 Plots of binomial and hypergeometric distributions

Definition 2.1.2. The set $\mathcal{D}(X)$ of all distributions over a set X can be defined as:

$$\mathcal{D}(X) \coloneqq \{\omega \colon X \to [0,1] \mid \mathit{supp}(\omega) \text{ is finite, and } \textstyle\sum_x \omega(x) = 1\}.$$

Such a function $\omega \colon X \to [0,1]$, with finite support and values adding up to one, is often called a *probability mass function*, abbreviated as pmf.

This $\mathcal{D}$ is functorial: for a function $f \colon X \to Y$ we have $\mathcal{D}(f) \colon \mathcal{D}(X) \to \mathcal{D}(Y)$ defined either as:

$$\mathcal{D}(f)\big(\textstyle\sum_i r_i|x_i\rangle\big) \coloneqq \sum_i r_i|f(x_i)\rangle \qquad \text{or as:} \qquad \mathcal{D}(f)(\omega)(y) \coloneqq \sum_{x \in f^{-1}(y)} \omega(x).$$

A distribution of the form $\mathcal{D}(f)(\omega) \in \mathcal{D}(Y)$, for $\omega \in \mathcal{D}(X)$, is sometimes called an *image distribution*. One also says that ω is pushed forward along f.

One has to check that $\mathcal{D}(f)(\omega)$ is a distribution again, that is, that its multi-

plicities add up to one. This works as follows.

$$\sum_{y \in Y} \mathcal{D}(f)(\omega)(y) = \sum_{y \in Y} \sum_{x \in f^{-1}(y)} \omega(x) = \sum_{x \in X} \omega(x) = 1.$$

We present two examples where functoriality of $\mathcal{D}$ is frequently used, but not always in explicit form.

Example 2.1.3. 1 Computing marginals of 'joint' distributions involves functoriality of $\mathcal{D}$. In general, one speaks of a *joint* distribution if its sample space is a product set, of the form $X_1 \times X_2$, or more generally, $X_1 \times \cdots \times X_n$, for $n \geq 2$. The i-th marginal of $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n)$ is defined as $\mathcal{D}(\pi_i)(\omega)$, via the i-th projection function $\pi_i \colon X_1 \times \cdots \times X_n \to X_i$.

For instance, the first marginal of the joint distribution,

$$\omega = \tfrac{1}{12}|H,0\rangle + \tfrac{1}{6}|H,1\rangle + \tfrac{1}{3}|H,2\rangle + \tfrac{1}{6}|T,0\rangle + \tfrac{1}{12}|T,1\rangle + \tfrac{1}{6}|T,2\rangle$$

on the product space $\{H,T\} \times \{0,1,2\}$ is obtained as:

$$\begin{aligned} \mathcal{D}(\pi_1)(\omega) &= \tfrac{1}{12}|\pi_1(H,0)\rangle + \tfrac{1}{6}|\pi_1(H,1)\rangle + \tfrac{1}{3}|\pi_1(H,2)\rangle \\ &\qquad + \tfrac{1}{6}|\pi_1(T,0)\rangle + \tfrac{1}{12}|\pi_1(T,1)\rangle + \tfrac{1}{6}|\pi_1(T,2)\rangle \\ &= \tfrac{1}{12}|H\rangle + \tfrac{1}{6}|H\rangle + \tfrac{1}{3}|H\rangle + \tfrac{1}{6}|T\rangle + \tfrac{1}{12}|T\rangle + \tfrac{1}{6}|T\rangle \\ &= \tfrac{7}{12}|H\rangle + \tfrac{5}{12}|T\rangle. \end{aligned}$$

2 Let $\omega \in \mathcal{D}(X)$ be distribution. In Chapter 4 we shall discuss random variables in detail, but here it suffices to know that it involves a function $R \colon X \to \mathbb{R}$ from the sample space of the distribution ω to the real numbers. Often, then, the notation

$$P[R = r] \in [0,1]$$

is used to indicate the probability that the random variable R takes value $r \in \mathbb{R}$.

Since we now know that $\mathcal{D}$ is functorial, we may apply it to the function $R \colon X \to \mathbb{R}$. It gives another function $\mathcal{D}(R) \colon \mathcal{D}(X) \to \mathcal{D}(\mathbb{R})$, so that $\mathcal{D}(R)(\omega)$ is a distribution on $\mathbb{R}$. We observe:

$$P[R = r] = \mathcal{D}(R)(\omega)(r) = \sum_{x \in R^{-1}(r)} \omega(x).$$

Thus, $P[R = (-)]$ is an image distribution, on $\mathbb{R}$. In the notation $P[R = r]$ the distribution ω on the sample space X is left implicit.

Here is a concrete example. Recall that we write $\mathit{pips} = \{1,2,3,4,5,6\}$ for the sample space of a dice. Let $M \colon \mathit{pips} \times \mathit{pips} \to \mathit{pips}$ take the maximum, so $M(i,j) = \max(i,j)$. We consider M as a function $\mathit{pips} \times \mathit{pips} \to \mathbb{R}$ via the

inclusion $\mathit{pips} \hookrightarrow \mathbb{R}$. Then, using the uniform distribution $\mathit{unif} \in \mathcal{D}(\mathit{pips} \times \mathit{pips})$,

$$\begin{aligned} P[M = k] &= \text{the probability that the maximum of two dice throws is } k \\ &= \mathcal{D}(M)(\mathit{unif})(k) \\ &= \sum_{i,j \text{ with } \max(i,j)=k} \mathit{unif}(i, j) \\ &= \sum_{i \leq k} \mathit{unif}(i, k) + \sum_{j<k} \mathit{unif}(k, j) \\ &= \frac{2k-1}{36}. \end{aligned}$$

In the notation of this book the image distribution $\mathcal{D}(M)(\mathit{unif})$ on pips is written as in the first line below.

$$\begin{aligned} \mathcal{D}(M)(\mathit{unif}) &= \tfrac{1}{36}|1\rangle + \tfrac{3}{36}|2\rangle + \tfrac{5}{36}|3\rangle + \tfrac{7}{36}|4\rangle + \tfrac{9}{36}|5\rangle + \tfrac{11}{36}|6\rangle \\ &= P[M = 1]\,\big|1\big\rangle + P[M = 2]\,\big|2\big\rangle + P[M = 3]\,\big|3\big\rangle \\ &\qquad + P[M = 4]\,\big|4\big\rangle + P[M = 5]\,\big|5\big\rangle + P[M = 6]\,\big|6\big\rangle. \end{aligned}$$

Since in our notation the underlying (uniform) distribution *unif* is explicit, we can also change it to another distribution $\omega \in \mathcal{D}(\mathit{pips} \times \mathit{pips})$ and still do the same computation. In fact, once we have seen product distributions in Section 2.4 we can compute $\mathcal{D}(M)(\omega_1 \otimes \omega_2)$ where ω_1 is a distribution for the first dice (say a uniform, fair one) and ω_2 a distribution for the second dice (which may be unfair). This notation in which states are written explicitly gives much flexibility in what we wish to express and compute, see also Subsection 4.1.1 below.

In the previous chapter we have seen that the sets $\mathcal{L}(X)$, $\mathcal{P}(X)$ and $\mathcal{M}(X)$ of lists, powersets and multisets all carry a monoid structure. Hence one may expect a similar result saying that $\mathcal{D}(X)$ forms a monoid, via an elementwise sum, like for multisets. But that does *not* work. Instead, one can take *convex sums* of distributions. This works as follows. Suppose we have two distributions $\omega, \rho \in \mathcal{D}(X)$ and a number $s \in [0, 1]$. Then we can form a new distribution $\sigma \in \mathcal{D}(X)$, as convex combination of ω and ρ, namely:

$$\sigma \coloneqq s \cdot \omega + (1 - s) \cdot \rho \qquad \text{that is} \qquad \sigma(x) = r \cdot \omega(x) + (1 - s) \cdot \rho(x). \tag{2.6}$$

This obviously generalises to an n-ary convex sum.

At this stage we shall not axiomatise structures with such convex sums; they are sometimes called simply 'convex sets' or 'barycentric algebras', see [155] or [67] for details. A brief historical account occurs in [99, Remark 2.9].

Remark 2.1.4. We have restricted ourselves to *finite* probability distributions, by requiring that the support of a distribution is a finite set. This is fine in many situations of practical interest, such as in Bayesian networks. But there are relevant distributions that have infinite support, like the *Poisson* distribution $\mathit{pois}[\lambda]$ on $\mathbb{N}$, with 'mean', 'rate' or 'intensity' parameter $\lambda \geq 0$. It can be described as infinite formal convex sum:

$$\mathit{pois}[\lambda] \coloneqq \sum_{k \in \mathbb{N}} e^{-\lambda} \cdot \frac{\lambda^k}{k!} \big| k \rangle. \tag{2.7}$$

This does not fit in $\mathcal{D}(\mathbb{N})$. Therefore we sometimes use $\mathcal{D}_\infty$ instead of $\mathcal{D}$, where the finiteness of support requirement has been dropped:

$$\mathcal{D}_\infty(X) \coloneqq \{\omega \colon X \to [0,1] \mid \textstyle\sum_x \omega(x) = 1\}. \tag{2.8}$$

Notice by the way that the multiplicities add up to one in the Poisson distribution because of the basic formula: $e^\lambda = \sum_{k \in \mathbb{N}} \frac{\lambda^k}{k!}$. The Poisson distribution is typically used for counts of rare events. The *rate* or *intensity* parameter λ is the average number of events per time period. The Poisson distribution then gives for each $k \in \mathbb{N}$ the probability of having k events per time period. This works when events occur independently.

Exercises 2.1.9 and 2.1.10 contain other examples of discrete distributions with infinite support, namely the geometric and negative-binomial distribution. Another illustration is the zeta (or zipf) distribution, see *e.g.* [145].

Exercises

2.1.1 Check that a marginal of a uniform distribution is again a uniform distribution; more precisely, $\mathcal{D}(\pi_1)(\mathit{unif}_{X \times Y}) = \mathit{unif}_X$.

2.1.2
1. Prove that $\mathit{flip} \colon [0,1] \to \mathcal{D}(2)$ is an isomorphism.
2. Check that $\mathit{flip}(r)$ is the same as $\mathit{bn}[1](r)$.
3. Describe the distribution $\mathit{bn}[3](\frac{1}{4})$ concretely and interepret this distribution in terms of coin flips.

2.1.3 We often write $n \coloneqq \{0, \ldots, n-1\}$ so that $0 = \emptyset$, $1 = \{0\}$ and $2 = \{0, 1\}$. Verify that:

$$\begin{array}{lllll} \mathcal{L}(0) \cong 1 & \mathcal{P}(0) \cong 1 & \mathcal{N}(0) \cong 1 & \mathcal{M}(0) \cong 1 & \mathcal{D}(0) \cong 0 \\ \mathcal{L}(1) \cong \mathbb{N} & \mathcal{P}(1) \cong 2 & \mathcal{N}(1) \cong \mathbb{N} & \mathcal{M}(1) \cong \mathbb{R}_{\geq 0} & \mathcal{D}(1) \cong 1 \\ & \mathcal{P}(n) \cong 2^n & \mathcal{N}(n) \cong \mathbb{N}^n & \mathcal{M}(n) \cong \mathbb{R}^n_{\geq 0} & \mathcal{D}(2) \cong [0,1]. \end{array}$$

The set $\mathcal{D}(n+1)$ is often called the n-simplex. Describe it as a subset of $\mathbb{R}^{n+1}$, and also as a subset of $\mathbb{R}^n$.

2.1.4 Let $X = \{a, b, c\}$ with draw $\varphi = 2|a\rangle + 3|b\rangle + 2|c\rangle \in \mathcal{N}[7](X)$.

1 Consider distribution $\omega = \frac{1}{3}|a\rangle + \frac{1}{2}|b\rangle + \frac{1}{6}|c\rangle$ and show that:

$$mn[7](\omega)(\varphi) \;=\; \tfrac{35}{432}.$$

Explain this outcome in terms of iterated single draws.

2 Consider urn $\psi = 4|a\rangle + 6|b\rangle + 2|c\rangle \in \mathcal{N}[12](X)$ and compute:

$$hg[7](\psi)(\varphi) \;=\; \tfrac{5}{33}.$$

2.1.5 Let a number $r \in [0, 1]$ and a finite set X be given. Show that:

$$\sum_{U \in \mathcal{P}(X)} r^{|U|} \cdot (1 - r)^{|X \setminus U|} \;=\; 1.$$

Can you generalise to numbers $r_1, \ldots, r_n$ and partitions of X of size n?

Hint: Recall the binomial and multinomial distribution from Example 2.1.1 (2) and recall Lemma 1.3.7.

2.1.6 Check that the powerbag operation from Exercise 1.6.8 can be turned into a probabilistic powerbag PPB via:

$$\mathsf{PPB}(\psi) \;\coloneqq\; \sum_{\varphi \leq \psi} \frac{\binom{\psi}{\varphi}}{2^{\|\psi\|}} \big|\varphi\rangle.$$

2.1.7 Let X be a non-empty finite set with n elements. Use Exercise 1.5.7 to check that, for the following *multiset distribution* yields a well-defined distribution on $\mathcal{N}[K](X)$.

$$mltdst[K]_X \;\coloneqq\; \sum_{\varphi \in \mathcal{N}[K](X)} \frac{(\varphi)}{n^K} \big|\varphi\rangle.$$

Describe $mltdst[4]_X$ for $X = \{a, b, c\}$.

2.1.8 1 Recall Theorem 1.5.2 (1) and conclude that $\lim_{n\to\infty} \binom{n+K}{K} \cdot r^n = 0$, for $r \in [0, 1)$. (This is general result: if partial sums of a series converge, the limit of the series itself is zero.)

2 Conclude that for $r \in (0, 1]$ one has:

$$\lim_{n\to\infty} bn[n+m](r)(m) \;=\; 0.$$

Explain yourself what this means.

2.1.9 For a (non-zero) probability $r \in (0, 1)$ one defines the *geometric distribution* $geo[r] \in \mathcal{D}_\infty(\mathbb{N})$ as:

$$geo[r] \;\coloneqq\; \sum_{k \in \mathbb{N}_{>0}} r \cdot (1 - r)^{k-1} \big|k\rangle.$$

It captures the probability of being successful for the first time after

$k-1$ unsuccesful tries. Prove that this is a distribution indeed: its multiplicities add up to one.

Hint: Recall the summation formula for geometric series, but don't confuse this geometric distribution with the hypergeometric distribution from Example 2.1.1 (3).

2.1.10 The negative binomial distribution is of the form $nbn[K](s) \in \mathcal{D}_\infty(\mathbb{N})$, for $K \geq 1$ and $s \in (0,1)$. It captures the probability of reaching K successes, with probability s, in $n+K$ trials.

$$\begin{aligned} nbn[K](s) &\coloneqq \sum_{n\geq 0} \left(\!\binom{K}{n}\!\right) \cdot s^K \cdot (1-s)^n \,\big|K+n\big\rangle \\ &= \sum_{n\geq 0} \binom{n+K-1}{K-1} \cdot s^K \cdot (1-s)^n \,\big|K+n\big\rangle \\ &= \sum_{m\geq K} \binom{m-1}{K-1} \cdot s^K \cdot (1-s)^{m-K} \,\big|m\big\rangle. \end{aligned}$$

Use Theorem 1.5.2 (or Exercise 1.5.9) to show that this forms a distribution. We shall describe negative multinomial distributions in Section **??**.

2.1.11 Prove that a distribution $\omega \in \mathcal{D}_\infty(X)$ necessarily has *countable* support.

Hint: Use that each set $\{x \in X \mid \omega(x) > \frac{1}{n}\}$, for $n > 0$, can have only finitely many elements.

2.1.12 Let $\omega \in \mathcal{D}(X)$ be an arbitrary distribution on a set X. We extend it to a distribution $\omega^\star$ on the set $\mathcal{L}(X)$ of lists of elements from X. We define the function $\omega^\star \colon \mathcal{L}(X) \to [0,1]$ by:

$$\omega^\star([x_1,\ldots,x_n]) \coloneqq \frac{\omega(x_1)\cdot\ldots\cdot\omega(x_n)}{2^{n+1}}.$$

1 Prove that $\omega^\star \in \mathcal{D}_\infty(\mathcal{L}(X))$.

2 Consider the function $f\colon \{a,b,c\} \to \{1,2\}$ with $f(a)=1$, $f(b)=1$, $f(c)=2$. Take $\omega = \frac{1}{3}|a\rangle + \frac{1}{4}|b\rangle + \frac{5}{12}|c\rangle \in \mathcal{D}(\{a,b,c\})$ and $\ell = [1,2,1] \in \mathcal{L}(\{1,2\})$. Show that:

$$\mathcal{D}(f)(\omega)^\star(\ell) = \tfrac{245}{27648} = \mathcal{D}_\infty(\mathcal{L}(f))(\omega^\star)(\ell).$$

Using the $\star$-operation as a function:

$$\mathcal{D}(X) \xrightarrow{(-)^\star} \mathcal{D}_\infty(\mathcal{L}(X))$$

we can describe the above equation as:

$$\begin{array}{ccc} \mathcal{D}(\{a,b,c\}) \ni \omega & \longmapsto & \omega^\star \in \mathcal{D}_\infty(\mathcal{L}(\{a,b,c\})) \\ \downarrow & & \downarrow \\ & & \mathcal{D}_\infty(\mathcal{L}(f))(\omega^\star) \\ & & \| \\ \mathcal{D}(\{1,2\}) \ni \mathcal{D}(f)(\omega) & \longmapsto & \mathcal{D}(f)(\omega)^\star \in \mathcal{D}_\infty(\mathcal{L}(\{1,2\})) \end{array}$$

3 Prove in general that $(-)^\star$ is a natural transformation from $\mathcal{D}$ to $\mathcal{D}_\infty \circ \mathcal{L}$.

2.2 Probabilistic channels

In the previous chapter we have seen that taking lists / powersets / multisets is functorial, giving functors written respectively as $\mathcal{L}$, $\mathcal{P}$, $\mathcal{M}$. Further, we have seen that these functors all have 'unit' and 'flatten' operations that satisfy standard equations that make $\mathcal{L}$, $\mathcal{P}$ and $\mathcal{M}$ into a monad. Moreover, in terms of these unit and flatten maps we have defined categories of channels $\mathbf{Chan}(\mathcal{L})$, $\mathbf{Chan}(\mathcal{P})$ and $\mathbf{Chan}(\mathcal{M})$. In this section we will show that the same monad structure exists for the distribution functor $\mathcal{D}$ and that we thus also have probabilistic channels, in a category $\mathbf{Chan}(\mathcal{D})$. In the remainder of this book the emphasis will be almost exclusively on this probabilistic case, so that 'channel' will standardly mean 'probabilistic channel'.

The unit and flatten maps for multisets from Subsection 1.4.2 restrict to distributions. The unit function $\mathit{unit} \colon X \to \mathcal{D}(X)$ is simply $\mathit{unit}(x) \coloneqq 1|x\rangle$. Flattening is the function $\mathit{flat} \colon \mathcal{D}(\mathcal{D}(X)) \to \mathcal{D}(X)$ with:

$$\mathit{flat}\Big(\textstyle\sum_i r_i|\omega_i\rangle\Big) \coloneqq \displaystyle\sum_{x\in X} \big(\textstyle\sum_i r_i \cdot \omega_i(x)\big)\big|x\big\rangle = \textstyle\sum_i r_i \cdot \omega_i.$$

The latter formulation uses a convex sum of distributions (2.6).

The only thing that needs to be checked is that flattening yields a convex sum, *i.e.* that its probabilities add up to one. This is easy:

$$\textstyle\sum_x \big(\sum_i r_i \cdot \omega_i(x)\big) = \sum_i r_i \cdot \sum_x \omega_i(x) = \sum_i r_i \cdot 1 = 1.$$

The familiar properties of unit and flatten hold for distributions too: the analogue of Lemma 1.4.3 holds, with 'multiset' replaced by 'distribution'. We conclude that $\mathcal{D}$, with these *unit* and *flat*, is a monad. The same holds for the 'infinite' variation $\mathcal{D}_\infty$ from Remark 2.1.4.

A probabilistic channel $c \colon X \nrightarrow Y$ is a function $c \colon X \to \mathcal{D}(Y)$. For a state / distribution $\omega \in \mathcal{D}(X)$ we can do state transformation and produce a new state

$c \gg\!= \omega \in \mathcal{D}(Y)$. This happens via the standard definition for $\gg\!=$ from Section 1.8, see especially (1.39) and (1.40):

$$c \gg\!= \omega \;:=\; \mathit{flat}\big(\mathcal{D}(c)(\omega)\big) \;=\; \sum_{y\in Y}\Big(\sum_{x\in X} c(x)(y)\cdot\omega(x)\Big)\big|y\rangle. \tag{2.9}$$

In the probabilistic case we get a probability distribution again, with probabilities adding up to one, since:

$$\sum_{y\in Y}\left(\sum_{x\in X} c(x)(y)\cdot\omega(x)\right) \;=\; \sum_{x\in X}\left(\sum_{y\in Y} c(x)(y)\right)\cdot\omega(x) \;=\; \sum_{x\in X} 1\cdot\omega(x) \;=\; 1.$$

Moreover, if we have another channel $d\colon Y \to \mathcal{D}(Z)$ we can form the composite channel:

$$d \mathbin{\odot} c \;:=\; \Big(X \xrightarrow{\;c\;} \mathcal{D}(Y) \xrightarrow{\mathcal{D}(d)} \mathcal{D}(\mathcal{D}(Z)) \xrightarrow{\mathit{flat}} \mathcal{D}(Z)\Big).$$

More explicitly, this gives $(d \mathbin{\odot} c)(x) = d \gg\!= c(x)$.

As a result we can form a category $\mathbf{Chan}(\mathcal{D})$ of probabilistic channels. Its objects are sets X and its morphisms $X \mathrel{-\!\!\circ\!\!\to} Y$ are probabilistic channels. They can be composed via $\odot$, see Lemma 1.8.3, with $\mathit{unit}\colon X \mathrel{-\!\!\circ\!\!\to} X$ as identity channel. Recall that an ordinary function $f\colon X \to Y$ can be turned into a (probabilistic) channel $\langle f\rangle\colon X \mathrel{-\!\!\circ\!\!\to} Y$. Explicitly, $\langle f\rangle(x) = 1|f(x)\rangle$. This can be formalised in terms of a functor $\mathbf{Sets} \to \mathbf{Chan}(\mathcal{D})$. Often, we don't write the $\langle -\rangle$ angles when the context makes clear what is meant.

We have already seen several examples of such probabilistic channels in Example 2.1.1, namely:

$$[0,1] \xrightarrow{\mathit{flip}} \{0,1\} \qquad \text{and} \qquad [0,1] \xrightarrow{\mathit{bn}[K]} \{0,1,\ldots,K\}.$$

The multinomial, hypergeometric, and Pólya distributions from (2.1), (2.3) and (2.4) can be organised as channels of the form:

$$\mathcal{D}(X) \xrightarrow{\mathit{mn}[K]} \mathcal{N}[K](X) \qquad \mathcal{N}[L](X) \xrightarrow{\mathit{hg}[K]} \mathcal{N}[K](X). \qquad \mathcal{N}_*(X) \xrightarrow{\mathit{pl}[K]} \mathcal{N}[K](X).$$

Example 2.2.1. We now present an example of a probabilistic channel, in the style of the earlier list/powerset/multiset channels in Example 1.8.2. We use again the sets $X = \{a,b,c\}$ and $Y = \{u,v\}$, with state $\omega = \frac{1}{6}|a\rangle + \frac{1}{2}|b\rangle + \frac{1}{3}|c\rangle \in \mathcal{D}(X)$ and channel $f\colon X \to \mathcal{D}(Y)$ given by:

$$f(a) \;=\; \tfrac{1}{2}|u\rangle + \tfrac{1}{2}|v\rangle \qquad f(b) \;=\; 1|u\rangle \qquad f(c) \;=\; \tfrac{3}{4}|u\rangle + \tfrac{1}{4}|v\rangle.$$

We then get as state transformation:

$$\begin{aligned} f \gg\!= \omega &= \mathit{flat}(\mathcal{D}(f)(\omega)) \\ &= \mathit{flat}(\tfrac{1}{6}|f(a)\rangle + \tfrac{1}{2}|f(b)\rangle + \tfrac{1}{3}|f(c)\rangle) \\ &= \mathit{flat}(\tfrac{1}{6}\big|\tfrac{1}{2}|u\rangle + \tfrac{1}{2}|v\rangle\big\rangle + \tfrac{1}{2}\big|1|u\rangle\big\rangle + \tfrac{1}{3}\big|\tfrac{3}{4}|u\rangle + \tfrac{1}{4}|v\rangle\big\rangle) \\ &= \tfrac{1}{12}|u\rangle + \tfrac{1}{12}|v\rangle + \tfrac{1}{2}|u\rangle + \tfrac{1}{4}|u\rangle + \tfrac{1}{12}|v\rangle \\ &= \tfrac{5}{6}|u\rangle + \tfrac{1}{6}|v\rangle. \end{aligned}$$

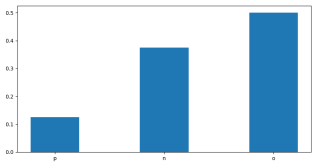
This state transformation $f \gg\!= \omega$ describes a mixture (convex sum) of the distributions $f(a), f(b), f(c)$, where the weights of the components of the mixture are given by the distribution ω, see also Exercise 2.2.3 (2). In practice we often compute transformation directly as convex sum of states, as in:

$$\begin{aligned} f \gg\!= \omega &= \tfrac{1}{6} \cdot \Big(\tfrac{1}{2}|u\rangle + \tfrac{1}{2}|v\rangle\Big) + \tfrac{1}{2} \cdot \Big(1|u\rangle\Big) + \tfrac{1}{3} \cdot \Big(\tfrac{3}{4}|u\rangle + \tfrac{1}{4}|v\rangle\Big) \\ &= \tfrac{5}{6}|u\rangle + \tfrac{1}{6}|v\rangle. \end{aligned}$$

This 'mixture' terminology is often used in a clustering context where the elements a, b, c are the components.

Example 2.2.2 (Copied from [79])**.** Let's assume we wish to capture the mood of a teacher, as a probabilistic mixture of three possible options namely: pessimistic (p), neutral (n) or optimistic (o). We thus have a three-element probability space $X = \{p, n, o\}$. We assume a mood distribution:

$$\sigma = \tfrac{1}{8}|p\rangle + \tfrac{3}{8}|n\rangle + \tfrac{1}{2}|o\rangle \qquad \text{with plot}$$



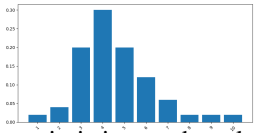
This mood thus tends towards optimism.

Associated with these different moods the teacher has different views on how pupils perform in a particular test. This performance is expressed in terms of marks, which can range from 1 to 10, where 10 is best. The probability space for these marks is written as $Y = \{1, 2, \ldots, 10\}$.

The view of the teacher is expressed via a channel $c\colon X \nrightarrow Y$. It is defined as via the following three mark distributions, one for each element in $X =$

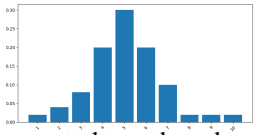
$\{p, n, o\}$.

$$\begin{aligned} c(p) \\ &= \tfrac{1}{50}|1\rangle + \tfrac{2}{50}|2\rangle + \tfrac{10}{50}|3\rangle + \tfrac{15}{50}|4\rangle + \tfrac{10}{50}|5\rangle \\ &\quad + \tfrac{6}{50}|6\rangle + \tfrac{3}{50}|7\rangle + \tfrac{1}{50}|8\rangle + \tfrac{1}{50}|9\rangle + \tfrac{1}{50}|10\rangle \end{aligned}$$



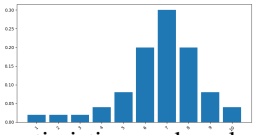
pessimistic mood marks

$$\begin{aligned} c(n) \\ &= \tfrac{1}{50}|1\rangle + \tfrac{2}{50}|2\rangle + \tfrac{4}{50}|3\rangle + \tfrac{10}{50}|4\rangle + \tfrac{15}{50}|5\rangle \\ &\quad + \tfrac{10}{50}|6\rangle + \tfrac{5}{50}|7\rangle + \tfrac{1}{50}|8\rangle + \tfrac{1}{50}|9\rangle + \tfrac{1}{50}|10\rangle \end{aligned}$$



neutral mood marks

$$\begin{aligned} c(o) \\ &= \tfrac{1}{50}|1\rangle + \tfrac{1}{50}|2\rangle + \tfrac{1}{50}|3\rangle + \tfrac{2}{50}|4\rangle + \tfrac{4}{50}|5\rangle \\ &\quad + \tfrac{10}{50}|6\rangle + \tfrac{15}{50}|7\rangle + \tfrac{10}{50}|8\rangle + \tfrac{4}{50}|9\rangle + \tfrac{2}{50}|10\rangle. \end{aligned}$$

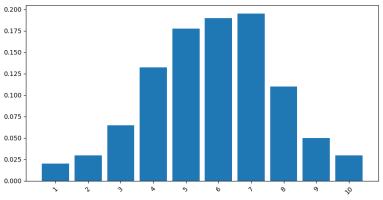


optimistic mood marks

Now that the state $\sigma \in \mathcal{D}(X)$ and the channel $c \colon X \nrightarrow Y$ have been described, we can form the transformed state $c \gg\!= \sigma \in \mathcal{D}(Y)$. Following the formulation (2.9) we get for each mark $i \in Y$ the probability:

$$\begin{aligned} (c \gg\!= \sigma)(i) &= \textstyle\sum_x \sigma(x) \cdot c(x)(i) \\ &= \sigma(p) \cdot c(p)(i) + \sigma(n) \cdot c(n)(i) + \sigma(o) \cdot c(o)(i). \end{aligned}$$

The outcome is in the plot below. It contains a convex combination of the above three distributions (and plots), for $c(p)$, $c(n)$ and $c(o)$, where the weights are determined by the mood distribution σ. This plot contains the 'predicted' marks, corresponding to the state of mind of the teacher.



This example will return in later chapters. There, the teacher will be confronted with the marks that the pupils actually obtain. This will lead the teacher to an

update of his/her mood. This situation forms an illustration of a predictive coding model, in which the human mind actively predicts the situation in the outside world, depending on its internal state — and updates it when confronted with the actual situation.

In the next example we show how products, multisets and distributions come together in an elementary combinatorial situation.

Example 2.2.3. Let X be an arbitrary set and K be a fixed positive natural number. The K-fold product $X^K = X \times \cdots \times X$ contains the lists of length K. As we have seen in (1.32), the accumulator function $acc\colon \mathcal{L}(X) \to \mathcal{N}(X)$ from (1.30) restricts to $acc\colon X^K \to \mathcal{N}[K](X)$, where, recall, $\mathcal{N}[K](X)$ is the set of multisets over X with K elements.

We ask ourselves: do these maps $acc\colon X^K \to \mathcal{N}[K](X)$ have inverses? Let $\varphi = \sum_i k_i|x_i\rangle \in \mathcal{N}[K](X)$ be a K-element natural multiset, so $\|\varphi\| = \sum_i k_i = K$, where we assume that i ranges over $1, 2, \ldots, n$. We can surely *choose* a list $\ell \in X^K$ with $acc(\ell) = \varphi$. All we have to do is put the elements in φ in a certain order. We can do so in many ways. How many? In Proposition 1.6.3 we have seen that the answer is given by the coefficient $\mathbf{(}\varphi\mathbf{)}$ of the multiset $\varphi \in \mathcal{N}[K](X)$, where:

$$\mathbf{(}\varphi\mathbf{)} = \frac{K!}{\prod_x \varphi(x)!} = \frac{K!}{k_1! \cdots k_n!} = \binom{K}{k_1, \ldots, k_n}.$$

Returning to our question about an inverse to $acc\colon X^K \to \mathcal{N}[K](X)$ we see that we can construct a *channel* in the reversed direction. We call it *arr* for 'arrange' and define it as $arr\colon \mathcal{N}[K](X) \nrightarrow X^K$ via uniform distributions:

$$arr(\varphi) \coloneqq \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{\mathbf{(}\varphi\mathbf{)}} \big|\vec{x}\big\rangle. \tag{2.10}$$

This is a sum over $\mathbf{(}\varphi\mathbf{)}$-many lists $\vec{x}$ with $acc(\vec{x}) = \varphi$. The size K of the multiset involved has disappeared from this formulation, so that we can view *arr* simply as a channel $arr\colon \mathcal{N}(X) \nrightarrow \mathcal{L}(X)$.

For instance, for $X = \{a, b\}$ with multiset $\varphi = 3|a\rangle + 1|b\rangle$ there are $\mathbf{(}\varphi\mathbf{)} = \binom{4}{3,1} = \frac{4!}{3!\cdot 1!} = 4$ arrangements of φ, namely $[a, a, a, b]$, $[a, a, b, a]$, $[a, b, a, a]$, and $[b, a, a, a]$, so that:

$$arr(3|a\rangle + 1|b\rangle) = \tfrac{1}{4}\big|a, a, a, b\big\rangle + \tfrac{1}{4}\big|a, a, b, a\big\rangle + \tfrac{1}{4}\big|a, b, a, a\big\rangle + \tfrac{1}{4}\big|b, a, a, a\big\rangle.$$

Our next question is: how are accumulator *acc* and arrangement *arr* related?

The following commuting triangle gives an answer.

$$\begin{array}{ccc} \mathcal{N}(X) & \xrightarrow{\;arr\;} & \mathcal{L}(X) \\ & \searrow^{unit} & \downarrow{\scriptstyle \langle acc \rangle} \\ & & \mathcal{N}(X) \end{array} \tag{2.11}$$

It combines a probabilistic channel $arr\colon \mathcal{N}(X) \to \mathcal{D}(\mathcal{L}(X))$ with an ordinary function $acc\colon \mathcal{L}(X) \to \mathcal{N}(X)$, promoted to a deterministic channel $\langle acc \rangle$. For the channel composition $\circ$ we make use of Lemma 1.8.3 (4):

$$\begin{aligned} (\langle acc \rangle \circ arr)(\varphi) &= \mathcal{D}(acc)(arr(\varphi)) \\ &= \mathcal{D}(acc)\left(\sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{\mathbf{(}\varphi\mathbf{)}} \big|\vec{x}\big\rangle\right) \\ &= \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{\mathbf{(}\varphi\mathbf{)}} \big|acc(\vec{x})\big\rangle \\ &= \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{\mathbf{(}\varphi\mathbf{)}} \big|\varphi\big\rangle = \mathbf{(}\varphi\mathbf{)} \cdot \frac{1}{\mathbf{(}\varphi\mathbf{)}} \big|\varphi\big\rangle = 1\big|\varphi\big\rangle = unit(\varphi). \end{aligned}$$

The above triangle (2.11) captures a very basic relationship between sequences, multisets and distributions, via the notion of (probabilistic) channel.

In the other direction, $arr(acc(\vec{x}))$ does not return $\vec{x}$. It yields a uniform distribution over all permutations of the sequence $\vec{x} \in X^K$.

Deleting and adding an element to a natural multiset are basic operations that are naturally described as channels.

Definition 2.2.4. For a set X and a number $K \in \mathbb{N}$ we define a draw-delete channel DD and also a draw-add channel DA in a situation:

$$\mathcal{N}[K](X) \underset{DA}{\overset{DD}{\rightleftarrows}} \mathcal{N}[K+1](X)$$

On $\psi \in \mathcal{N}[K+1](X)$ and $\varphi \in \mathcal{N}[K](X)$ they are defined as:

$$\begin{aligned} DD(\psi) &:= \sum_{x \in supp(\psi)} \frac{\psi(x)}{K+1} \big|\psi - 1|x\big\rangle\big\rangle \\ DA(\chi) &:= \sum_{x \in supp(\chi)} \frac{\chi(x)}{K} \big|\chi + 1|x\big\rangle\big\rangle \end{aligned} \tag{2.12}$$

For DA we need to assume that χ is non-empty, or equivalently, that $K > 0$.

In the draw-delete case DD one draws single elements $x \in X$ from the urn ψ, with probability determined by the number of occurrences $\psi(x) \in \mathbb{N}$ of x

in the urn ψ. This drawn element x is subsequently removed from the urn via subtraction $\psi - 1|x\rangle$.

In the draw-add case DA one also draws single elements x from the urn χ, but instead of deleting x one adds an extra copy of x, via the sum $\chi + 1|x\rangle$. This is typical for Pólya style urns, see [115] or Section 3.4 for more information.

These channels are not each other's inverses. For instance:

$$\begin{aligned} DD(3|a\rangle + 1|b\rangle) &= \tfrac{3}{4}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|3|a\rangle\big\rangle \\ DA(2|a\rangle + 1|b\rangle) &= \tfrac{2}{3}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{3}\big|2|a\rangle + 2|b\rangle\big\rangle. \end{aligned}$$

In a next step we see that neither $DA \circ DD$ nor $DD \circ DA$ is the identity.

$$\begin{aligned} &(DA \circ DD)(3|a\rangle + 1|b\rangle) \\ &= DA \gg\!= \Big(\tfrac{3}{4}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|3|a\rangle\big\rangle\Big) \\ &= \tfrac{3}{4}\Big(\tfrac{2}{3}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{3}\big|2|a\rangle + 2|b\rangle\big\rangle\Big) + \tfrac{1}{4}\Big(1\big|4|a\rangle\big\rangle\Big) \\ &= \tfrac{1}{2}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|2|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{4}\big|4|a\rangle\big\rangle \\ &(DD \circ DA)(2|a\rangle + 1|b\rangle) \\ &= DD \gg\!= \Big(\tfrac{2}{3}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{3}\big|2|a\rangle + 2|b\rangle\big\rangle\Big) \\ &= \tfrac{2}{3}\Big(\tfrac{3}{4}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|3|a\rangle\big\rangle\Big) + \tfrac{1}{3}\Big(\tfrac{1}{2}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{2}\big|2|a\rangle + 1|b\rangle\big\rangle\Big) \\ &= \tfrac{1}{2}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{6}\big|3|a\rangle\big\rangle + \tfrac{1}{6}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{6}\big|2|a\rangle + 1|b\rangle\big\rangle. \end{aligned}$$

When we iterate the draw-and-delete and the draw-and-add channels we obtain distributions that strongly remind us of the hypergeometric and Pólya distribution, see Example 2.1.1 (3) and (4). But they are not the same. The iterations below describe not what is drawn from the urn — as in the hypergeometric and Pólya cases — but what is left in the urn after such draws. The full picture appears in Chapter 3.

The iteration of draws is intuitively very natural, leading to subtraction and addition of the draws φ, in the formulas below. The important thing to note is that the mathematical formalisation of this intuition works because these draws are channels and are composed via channel composition $\circ$.

Theorem 2.2.5. *Iterating $K \in \mathbb{N}$ times the draw-and-delete and draw-and-add channels from Definition 2.2.4 yields channels:*

$$\mathcal{N}[L](X) \underset{DA^K}{\overset{DD^K}{\rightleftarrows}} \mathcal{N}[L+K](X)$$

On $\psi \in \mathcal{N}[L+K](X)$ and $\chi \in \mathcal{N}[L](X)$ they satisfy:

$$DD^K(\psi) = \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L+K}{K}} \big|\psi - \varphi\big\rangle \qquad DA^K(\chi) = \sum_{\varphi \leq_K \chi} \frac{\binom{\binom{\chi}{\varphi}}{}}{\binom{\binom{L}{K}}{}} \big|\chi + \varphi\big\rangle.$$

The probabilities in these expressions add up to one by Lemma 1.6.2 and Proposition 1.7.3.

Proof. For both equations we use induction on $K \in \mathbb{N}$. In both cases the only option for φ in $\mathcal{N}[0](X)$ is the empty multiset $\mathbf{0}$, so that $DD^0(\psi) = 1|\psi\rangle$ and $DA^0(\chi) = 1|\chi\rangle$.

For the induction steps we make extensive use of the equations in Exercise 1.7.6. In those cases we shall put 'E' above the equation. We start with the induction step for draw-delete. For $\psi \in \mathcal{N}[L+K+1](X)$,

$$\begin{aligned}
DD^{K+1}(\psi) &= DD^K \gg\!= DD(\psi) \\
&\overset{(2.12)}{=} DD^K \gg\!= \left(\sum_{x \in supp(\psi)} \frac{\psi(x)}{L+K+1} \big|\psi - 1|x\rangle\big\rangle \right) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in supp(\psi)} \frac{\psi(x)}{L+K+1} \cdot DD^K(\psi - 1|x\rangle)(\varphi) \big|\varphi\big\rangle \\
&\overset{(IH)}{=} \sum_{x \in supp(\psi)} \sum_{\chi \leq_K \psi - 1|x\rangle} \frac{\psi(x)}{L+K+1} \cdot \frac{\binom{\psi - 1|x\rangle}{\chi}}{\binom{L+K}{K}} \big|\psi - 1|x\rangle - \chi\big\rangle \\
&\overset{(E)}{=} \sum_{x \in supp(\psi)} \sum_{\chi \leq_K \psi - 1|x\rangle} \frac{\chi(x)+1}{K+1} \cdot \frac{\binom{\psi}{\chi+1|x\rangle}}{\binom{L+K+1}{K+1}} \big|\psi - (\chi + 1|x\rangle)\big\rangle \\
&= \sum_{\varphi \leq_{K+1} \psi} \sum_{x \in supp(\psi)} \frac{\varphi(x)}{K+1} \cdot \frac{\binom{\psi}{\varphi}}{\binom{L+K+1}{K+1}} \big|\psi - \varphi\big\rangle \\
&= \sum_{\varphi \leq_{K+1} \psi} \frac{\binom{\psi}{\varphi}}{\binom{L+K+1}{K+1}} \big|\psi - \varphi\big\rangle, \qquad \text{since } \|\varphi\| = K+1.
\end{aligned}$$

We reason basically in the same way for draw-add, and now also use Exer-

cise 1.7.5. For $\chi \in \mathcal{N}[L](X)$,

$$\begin{aligned}
DA^{K+1}(\chi) &\overset{(2.12)}{=} DA^K \gg\!= \left(\sum_{x \in supp(\chi)} \frac{\chi(x)}{L} \big|\chi + 1|x\rangle\big\rangle \right) \\
&\overset{(IH)}{=} \sum_{x \in supp(\chi)} \sum_{\psi \leq_K \chi + 1|x\rangle} \frac{\chi(x)}{L} \cdot \frac{\binom{\chi+1|x\rangle}{\psi}}{\binom{L+1}{K}} \big|\chi + 1|x\rangle + \psi\big\rangle \\
&\overset{(E)}{=} \sum_{x \in supp(\chi)} \sum_{\psi \leq_K \chi + 1|x\rangle} \frac{\psi(x) + 1}{K+1} \cdot \frac{\binom{\chi}{\psi+1|x\rangle}}{\binom{L}{K+1}} \big|\chi + \psi + 1|x\rangle\big\rangle \\
&= \sum_{\varphi \leq_{K+1} \chi} \sum_{x \in supp(\chi)} \frac{\varphi(x)}{K+1} \cdot \frac{\binom{\chi}{\varphi}}{\binom{L}{K+1}} \big|\chi + \varphi\big\rangle \\
&= \sum_{\varphi \leq_{K+1} \chi} \sum_{x \in supp(\chi)} \frac{\binom{\chi}{\varphi}}{\binom{L}{K+1}} \big|\chi + \varphi\big\rangle. \qquad \text{❑}
\end{aligned}$$

Exercises

2.2.1 Consider the $\mathcal{D}$-channel $f\colon \{a,b,c\} \nrightarrow \{u,v\}$ from Example 2.2.1, together with a new $\mathcal{D}$-channel $g\colon : \{u,v\} \nrightarrow \{1,2,3,4\}$ given by:

$$g(u) = \tfrac{1}{4}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{2}|3\rangle + \tfrac{1}{8}|4\rangle \qquad g(v) = \tfrac{1}{4}|1\rangle + \tfrac{1}{8}|3\rangle + \tfrac{5}{8}|3\rangle.$$

Describe the general $g \circ f\colon \{a,b,c\} \nrightarrow \{1,2,3,4\}$ concretely.

2.2.2 Formulate and prove the analogue of Lemma 1.4.3 for $\mathcal{D}$, instead of for $\mathcal{M}$.

2.2.3 Notice that for a probabilistic channel c one can describe state transformation as a (convex) sum/mixture of states $c(x)$: that is, as:

$$c \gg\!= \omega = \sum\nolimits_x \omega(x) \cdot c(x).$$

2.2.4 Identify the channel f and the state ω in Example 2.2.1 with matrices:

$$M_f = \begin{pmatrix} \frac{1}{2} & 1 & \frac{3}{4} \\ \frac{1}{2} & 0 & \frac{1}{4} \end{pmatrix} \qquad\qquad M_\omega = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{2} \\ \frac{1}{3} \end{pmatrix}$$

Such matrices are called stochastic, since each of their columns has non-negative entries that add up to one.

Check that the matrix associated with the transformed state $f \gg\!= \omega$ is the matrix-column multiplication $M_f \cdot M_\omega$.

(A general description appears in Remark 4.3.5.)

2.2.5 Prove that state transformation along $\mathcal{D}$-channels preserves convex combinations (2.6), that is, for $r \in [0,1]$,

$$f \gg\!= (r \cdot \omega + (1-r) \cdot \rho) \;=\; r \cdot (f \gg\!= \omega) + (1-r) \cdot (f \gg\!= \rho).$$

2.2.6 Write $\pi \colon X^{K+1} \to X^K$ for the obvious projection function. Show that the following diagram of channels commutes.

$$\begin{array}{ccc} \mathcal{N}[K+1](X) & \xrightarrow{\;DD\;} & \mathcal{N}[K](X) \\ {\scriptstyle arr}\downarrow & & \downarrow{\scriptstyle arr} \\ X^{K+1} & \xrightarrow{\;\pi\;} & X^X \end{array}$$

2.2.7 Consider the arrangement channels $arr \colon \mathcal{N}(X) \nrightarrow \mathcal{L}(X)$ from Example 2.2.3. Prove that arr is natural: for each function $f \colon X \to Y$ the following diagram commutes.

$$\begin{array}{ccc} \mathcal{N}(X) & \xrightarrow{\;arr\;} & \mathcal{D}(\mathcal{L}(X)) \\ {\scriptstyle \mathcal{N}(f)}\downarrow & & \downarrow{\scriptstyle \mathcal{D}(\mathcal{L}(f))} \\ \mathcal{N}(Y) & \xrightarrow{\;arr\;} & \mathcal{D}(\mathcal{L}(Y)) \end{array}$$

(This is far from easy; you may wish to check first what this means in a simple case and then content yourself with a 'proof by example'.)

2.2.8 Show that the draw-and-delete and draw-and-add maps DD and DA in Definition 2.2.4 are natural, from $\mathcal{N}[K+1]$ to $\mathcal{D} \circ \mathcal{N}[K]$, and from $\mathcal{N}[K]$ to $\mathcal{D} \circ \mathcal{N}[K+1]$.

2.2.9 Recall from (1.34) that the set $\mathcal{N}[K](X)$ of K-sized natural multisets over a set X with n has contains $\binom{\binom{n}{K}}{}$ members. We write $unif_K \in \mathcal{D}(\mathcal{N}[K](X))$ for the uniform distribution over such multisets.

1 Check that:

$$unif_K \;=\; \sum_{\varphi \in \mathcal{N}[K](n)} \frac{1}{\binom{\binom{n}{K}}{}} \big|\varphi\big\rangle.$$

2 Prove that these $unif_K$ distributions match the chain of draw-and-delete maps $DD \colon \mathcal{N}[K+1](X) \nrightarrow \mathcal{N}[K](X)$ in the sense that:

$$DD \gg\!= unif_{K+1} \;=\; unif_K.$$

In categorical language this means that these uniform distributions form a cone.

3 Check that the draw-add maps do *not* preserve uniform distributions. Check for instance that for $X = \{a, b\}$,

$$\begin{aligned} DA \gg\!= unif_3 \;=\; \tfrac{1}{4}\big|4|a\rangle\big\rangle + \tfrac{1}{6}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{6}\big|2|a\rangle + 2|b\rangle\big\rangle \\ + \tfrac{1}{6}\big|1|a\rangle + 3|b\rangle\big\rangle + \tfrac{1}{4}\big|4|b\rangle\big\rangle. \end{aligned}$$

2.2.10 Let X be a finite set, say with n elements, of the form $X = \{x_1, \ldots, x_n\}$. Define for $K \geq 1$,

$$\sigma_K \coloneqq \sum_{1 \leq i \leq n} \tfrac{1}{n} \big|K|x_i\rangle\big\rangle \in \mathcal{D}(\mathcal{N}[K](X)).$$

Thus:

$$\begin{aligned} \sigma_1 &= \tfrac{1}{n}\big|1|x_1\rangle\big\rangle + \cdots + \tfrac{1}{n}\big|1|x_n\rangle\big\rangle \\ \sigma_2 &= \tfrac{1}{n}\big|2|x_1\rangle\big\rangle + \cdots + \tfrac{1}{n}\big|2|x_n\rangle\big\rangle, \qquad \textit{etc.} \end{aligned}$$

Show that these σ_K form a cone, both for draw-delete and for draw-add:

$$DD \gg\!= \sigma_{K+1} = \sigma_K \qquad \text{and} \qquad DA \gg\!= \sigma_K = \sigma_{K+1}.$$

Thus, the whole sequence $(\sigma_K)_{K \in \mathbb{N}_{>0}}$ can be generated from $\sigma_1 = \mathit{unif}_{\mathcal{N}[1](X)}$ by repeated application of DA.

2.2.11 Recall the bijective correspondences from Exercise 1.8.4.

1 Let X, Y be finite sets and $c \colon X \to \mathcal{D}(Y)$ be a $\mathcal{D}$-channel. We can then define an $\mathcal{M}$-channel $c^\dagger \colon Y \to \mathcal{M}(X)$ by swapping arguments: $c^\dagger(y)(x) = c(x)(y)$. We call c a 'bi-channel' if $c^\dagger$ is also a $\mathcal{D}$-channel, *i.e.* if $\sum_x c(x)(y) = 1$ for each $y \in Y$.
Prove that the identity channel is a bi-channel and that bi-channels are closed under composition.

2 Take $A = \{a_0, a_1\}$ and $B = \{b_0, b_1\}$ and define a channel $\mathit{bell} \colon A \times 2 \to \mathcal{D}(B \times 2)$ as:

$$\begin{aligned} \mathit{bell}(a_0, 0) &= \tfrac{1}{2}|b_0, 0\rangle + \tfrac{3}{8}|b_1, 0\rangle + \tfrac{1}{8}|b_1, 1\rangle \\ \mathit{bell}(a_0, 1) &= \tfrac{1}{2}|b_0, 1\rangle + \tfrac{1}{8}|b_1, 0\rangle + \tfrac{3}{8}|b_1, 1\rangle \\ \mathit{bell}(a_1, 0) &= \tfrac{3}{8}|b_0, 0\rangle + \tfrac{1}{8}|b_0, 1\rangle + \tfrac{1}{8}|b_1, 0\rangle + \tfrac{3}{8}|b_1, 1\rangle \\ \mathit{bell}(a_1, 1) &= \tfrac{1}{8}|b_0, 0\rangle + \tfrac{3}{8}|b_0, 1\rangle + \tfrac{3}{8}|b_1, 0\rangle + \tfrac{1}{8}|b_1, 1\rangle \end{aligned}$$

Check that bell is a bi-channel.
(It captures the famous Bell table from quantum theory; we have deliberately used open spaces in the above description of the channel bell so that non-zero entries align, giving a 'bi-stochastic' matrix, from which one can read $\mathit{bell}^\dagger$ vertically.)

2.2.12 Check that the inclusions $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ form a map of monads, as described in Definition 1.9.2.

2.2.13 Recall from Exercise 1.9.6 that for a fixed set A the mapping $X \mapsto X + A$ is a monad. Prove that $X \mapsto \mathcal{D}(X + A)$ is also a monad.

(The latter monad will be used in Chapter **??** to describe Markov models with outputs. A composition of two monads is not necessarily

again a monad, but it is when there is a so-called distributive law between the monads, see *e.g.* [70, Chap. 5] for details.)

2.3 Frequentist learning: from multisets to distributions

We have introduced distributions as special multisets, namely as multisets in which the multiplicities add up to one, so that $\mathcal{D}(X) \subseteq \mathcal{M}(X)$. There are many other relations and interactions between distributions and multisets. As we have mentioned, an urn containing coloured balls can be described aptly as a multiset. The distribution for drawing a ball can then be derived from the multiset. In this section we shall describing this situation in terms of a mapping from multisets to distributions, called 'frequentist learning' since it basically involves counting. Later on we shall see that other forms of learning from data can be described in terms of passages from multisets to distributions. This forms a central topic.

In earlier sections we have seen several (natural) mappings between collection types, in the form of support maps, see the overview in Diagram (1.30). We are now adding the mapping $\mathit{Flrn} \colon \mathcal{M}_*(X) \to \mathcal{D}(X)$, see [76]. The name *Flrn* stands for 'frequentist learning', and may be pronounced as 'eff-learn'. The frequentist interpretation of probability theory views probabilities as long term frequencies of occurrences. Here, these occurrences are given via multisets, which form the inputs of the *Flrn* function. Later on, in Theorem 4.5.7 we show that these outcomes of frequentist learning ly dense in the set of distributions, which means that we can approximate each distribution with arbitrary precision via frequentist learning of (natural) multisets.

Recall that $\mathcal{M}_*(X)$ is the collection of *non-empty* multisets $\sum_i r_i|x_i\rangle$, with $r_i \neq 0$ for at least one index i. Equivalently one can require that the sum $s \coloneqq \sum_i r_i = \| \sum_i r_i|x_i\rangle\|$ is non-zero.

The *Flrn* maps turns a (non-empty) multiset into a distribution, essentially by normalisation:

$$\mathit{Flrn}\Big(r_1\big|x_1\big\rangle + \cdots + r_k\big|x_k\big\rangle\Big) \coloneqq \frac{r_1}{s}\big|x_1\big\rangle + \cdots + \frac{r_k}{s}\big|x_k\big\rangle \quad \text{where } s \coloneqq \textstyle\sum_i r_i. \tag{2.13}$$

The normalisation step forces the sum on the right-hand side to be a convex sum, with factors adding up to one. Clearly, from an empty multiset we cannot learn a distribution — technically because the above sum s is then zero so that we cannot divide by s.

Using scalar multiplication from Lemma 1.4.2 (2) we can define the *Flrn*

function more succintly:

$$\mathit{Flrn}(\varphi) \coloneqq \frac{1}{\|\varphi\|} \cdot \varphi \qquad \text{where} \qquad \|\varphi\| \coloneqq \textstyle\sum_x \varphi(x). \tag{2.14}$$

Example 2.3.1. We present two illustrations of frequentist learning.

1 Suppose we have some coin of which the bias is unkown. There are experimental data showing that after tossing the coin 50 times, 20 times come up head (H) and 30 times yield tail (T). We can present these data as a *multiset* $\varphi = 20|H\rangle + 30|T\rangle \in \mathcal{M}_*(\{H, T\})$. When we wish to learn the resulting probabilities, we apply the frequentist learning map *Flrn* and get a *distribution* in $\mathcal{D}(\{H, T\}$, namely:

$$\mathit{Flrn}(\varphi) \;=\; \tfrac{20}{20+30}|H\rangle + \tfrac{30}{20+30}|T\rangle \;=\; \tfrac{2}{5}|H\rangle + \tfrac{3}{5}|T\rangle.$$

Thus, the bias (twowards head) is $\frac{2}{5}$. In this simple case we could have obtained this bias immediately from the data, but the *Flrn* map captures the general mechanism.

Notice that with frequentist learning, more (or less) consistent data gives the same outcome. For instance if we knew that 40 out of 100 tosses were head, or 2 out of 5, we would still get the same bias. Intuitively, these data give more (or less) confidence about the data. These aspects are not covered by frequentist learning, but by a more sophisticated form of 'Bayesian' learning. Another disadvantage of the rather primitive form of frequentist learning is that prior knowledge, if any, about the bias is not taken into account.

2 Recall the medical table (1.18) captured by the multiset $\tau \in \mathcal{N}(B \times M)$. Learning from τ yields the following joint distribution:

$$\begin{aligned}\mathit{Flrn}(\tau) \;=\; & 0.1|H, 0\rangle + 0.35|H, 1\rangle + 0.25|H, 2\rangle \\ & + 0.05|L, 0\rangle + 0.1|L, 1\rangle + 0.15|L, 2\rangle.\end{aligned}$$

Such a distribution, directly derived from a table, is sometimes called an *empirical* distribution [33].

In the above coin example we saw a property that is typical of frequentist learning, namely that learning from more of the same does not have any effect. We can make this precise via the equation:

$$\mathit{Flrn}(K \cdot \varphi) \;=\; \mathit{Flrn}(\varphi) \qquad \text{for } K > 0. \tag{2.15}$$

In [57] it is argued that in general, people are not very good at probabilistic (esp. Bayesian) reasoning, but that they are much better at reasoning with "frequency formats". To put it simply: the information (from $\mathcal{D}$) that there is a 0.04

probability of getting a disease is more difficult to process than the information (from $\mathcal{M}$) that 4 out of 100 people get the disease. In the current setting these frequency formats would correspond to natural multisets; they can be turned into distributions via the frequentist learning map *Flrn*. In the sequel we regularly return to the interaction between multisets and distributions in relation to drawing from an urn (in Chapter 3) and to learning (in Chapter **??**).

It turns out that the learning map *Flrn* is 'natural', in the sense that it works uniformly for each set.

Lemma 2.3.2. *The frequentist learning maps* $\mathrm{Flrn}\colon \mathcal{M}_*(X) \to \mathcal{D}(X)$ *from* (2.13) *are natural in* X*. This means that for each function* $f\colon X \to Y$ *the following diagram commutes.*

$$\begin{array}{ccc} \mathcal{M}_*(X) & \xrightarrow{\mathit{Flrn}} & \mathcal{D}(X) \\ {\scriptstyle \mathcal{M}_*(f)}\downarrow & & \downarrow{\scriptstyle \mathcal{D}(f)} \\ \mathcal{M}_*(Y) & \xrightarrow[\mathit{Flrn}]{} & \mathcal{D}(Y) \end{array}$$

As a result, frequentist learning commutes with marginalisation (via projection functions), see also Subsection 2.5.1.

Proof. Pick an arbitrary non-empty multiset $\varphi = \sum_i r_i|x_i\rangle$ in $\mathcal{M}_*(X)$ and write $s := \|\varphi\| = \sum_i r_i$. By non-emptyness of φ we have $s \neq 0$. Then:

$$\begin{aligned} (\mathit{Flrn} \circ \mathcal{M}_*(f))(\varphi) &= \mathit{Flrn}(\textstyle\sum_i r_i|f(x_i)\rangle) \\ &= \textstyle\sum_i \frac{r_i}{s}|f(x_i)\rangle) \\ &= \mathcal{D}(f)(\textstyle\sum_i \frac{r_i}{s}|x_i\rangle) \\ &= (\mathcal{D}(f) \circ \mathit{Flrn})(\varphi). \end{aligned}$$ ❑

We can apply this basic result to the medical data in Table (1.18), via the multiset $\tau \in \mathcal{N}(B \times M)$. We have already seen in Section 1.4 that the multiset-marginals $\mathcal{N}(\pi_i)(\tau)$ produce the marginal columns and rows, with their totals. We can learn the distributions from the colums as:

$$\mathit{Flrn}(\mathcal{M}(\pi_1)(\tau)) = \mathit{Flrn}(70|H\rangle + 30|L\rangle) = 0.7|H\rangle + 0.3|L\rangle.$$

We can also take the distribution-marginal of the 'learned' distribution from the table, as described in Example 2.3.1 (2):

$$\begin{aligned} \mathcal{M}(\pi_1)(\mathit{Flrn}(\tau)) &= (0.1 + 0.35 + 0.25)|H\rangle + (0.05 + 0.1 + 0.15)|L\rangle \\ &= 0.7|H\rangle + 0.3|L\rangle. \end{aligned}$$

Hence the basic operations of learning and marginalisations commute. This is a simple result, which many practitioners in probability are surely aware

of, at an intuitive level, but maybe not in the mathematically precise form of Lemma 2.3.2.

Drawing an object from an urn is an elementary operation in probability theory, which involves frequentist learning *Flrn*. For instance, the draw-and-delete and draw-and-add operations *DD* and *DA* from Definition 2.2.4 can be described, for urns ψ and χ with $\|\psi\| = K + 1$ and $\|\chi\| = K$, as:

$$DD(\psi) = \sum_{x \in supp(\psi)} \frac{\psi(x)}{K+1} \big|\psi - 1|x\rangle\big\rangle = \sum_{x \in supp(\psi)} Flrn(\psi)(x) \big|\psi - 1|x\rangle\big\rangle$$

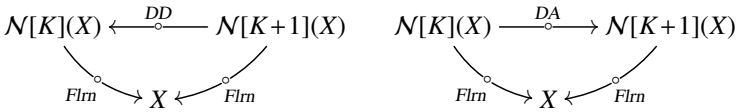
$$DA(\chi) = \sum_{x \in supp(\chi)} \frac{\chi(x)}{K} \big|\chi + 1|x\rangle\big\rangle = \sum_{x \in supp(\chi)} Flrn(\chi)(x) \big|\chi + 1|x\rangle\big\rangle.$$

Since this drawing takes the multiplicities into account, the urns after the *DD* and *DA* draws have the same frequentist distribution as before, but only if we interpret "after" as channel composition $\circ$. This is the content of the following basic result.

Theorem 2.3.3. *One has both:*

$$Flrn \circ DD = Flrn \qquad \text{and} \qquad Flrn \circ DA = Flrn.$$

Equivalently, the following two triangles of channels commute.



Proof. For the proof of commutation of draw-delete and frequentist learning, we take $\psi \in \mathcal{M}[K+1](X)$ and $y \in X$. Then:

$$\begin{aligned}
(Flrn \circ DD)(\psi)(y) &= \sum_{\varphi \in \mathcal{N}[K](X)} DD(\psi)(\varphi) \cdot Flrn(\varphi)(y) \\
&\overset{(2.12)}{=} \sum_{x \in X} \frac{\psi(x)}{K+1} \cdot Flrn(\psi - 1|x\rangle)(y) \\
&= \frac{\psi(y)}{K+1} \cdot \frac{\psi(y) - 1}{K} + \sum_{x \neq y} \frac{\psi(x)}{K+1} \cdot \frac{\psi(y)}{K} \\
&= \frac{\psi(y)}{K(K+1)} \cdot \left(\psi(y) - 1 + \sum_{x \neq y} \psi(x)\right) \\
&= \frac{\psi(y)}{K(K+1)} \cdot \left(\left(\sum_{x} \psi(x)\right) - 1\right) \\
&= \frac{\psi(y)}{K(K+1)} \cdot ((K+1) - 1) = \frac{\psi(y)}{K+1} = Flrn(\psi)(y).
\end{aligned}$$

Similarly, for $\chi \in \mathcal{N}[K](X)$, where $K > 0$, and $y \in X$,

$$\begin{aligned}
(\mathit{Flrn} \circ \mathit{DA})(\chi)(y) &\overset{(2.12)}{=} \sum_{x \in X} \mathit{Flrn}(\chi + 1|x\rangle)(y) \cdot \frac{\chi(x)}{K} \\
&= \frac{\chi(y)+1}{K+1} \cdot \frac{\chi(y)}{K} + \sum_{x \neq y} \frac{\chi(y)}{K+1} \cdot \frac{\chi(x)}{K} \\
&= \frac{\chi(y)}{K(K+1)} + \frac{\chi(y)}{K+1} \cdot \left(\frac{\chi(y)}{K} + \sum_{x \neq y} \frac{\chi(x)}{K} \right) \\
&= \frac{\chi(y)}{K(K+1)} + \frac{\chi(y)}{K+1} \cdot \sum_{x} \frac{\chi(x)}{K} \\
&= \frac{\chi(y)}{K(K+1)} + \frac{\chi(y)}{K+1} \\
&= \frac{\chi(y) + K \cdot \chi(y)}{K(K+1)} = \frac{\chi(y)}{K} = \mathit{Flrn}(\chi)(y). \qquad \text{❑}
\end{aligned}$$

Remark 2.3.4. In Lemma 2.3.2 we have seen that $\mathit{Flrn} \colon \mathcal{M}_* \to \mathcal{D}$ is a natural transformation. Since both $\mathcal{M}_*$ and $\mathcal{D}$ are monads, one can ask if *Flrn* is also a map of monads. This would mean that *Flrn* also commutes with the unit and flatten maps, see Definition 1.9.2. This is *not* the case.

It is easy to see that *Flrn* commutes with the unit maps, simply because $\mathit{Flrn}(1|x\rangle) = 1|x\rangle$. But commutation with flatten's fails. Here is a simple counterexample. Consider the multiset of multsets $\Phi \in \mathcal{M}(\mathcal{M}(\{a, b, c\}))$ given by:

$$\Phi \coloneqq 1\Big| 2|a\rangle + 4|c\rangle \Big\rangle + 2\Big| 1|a\rangle + 1|b\rangle + 1|c\rangle \Big\rangle.$$

First taking (multiset) multiplication, and then doing frequentist learning gives:

$$\mathit{Flrn}\Big(\mathit{flat}(\Phi)\Big) = \mathit{Flrn}\Big(4|a\rangle + 2|b\rangle + 6|c\rangle\Big) = \tfrac{1}{3}|a\rangle + \tfrac{1}{6}|b\rangle + \tfrac{1}{2}|c\rangle.$$

However, first (outer en inner) learning and then doing (distribution) multiplication yields:

$$\begin{aligned}
\mathit{flat}\Big(\mathit{Flrn}(\mathcal{M}(\mathit{Flrn})(\Phi))\Big) &= \tfrac{1}{3}\Big| \tfrac{1}{3}|a\rangle + \tfrac{2}{3}|c\rangle \Big\rangle + \tfrac{2}{3}\Big| \tfrac{1}{3}|a\rangle + \tfrac{1}{3}|b\rangle + \tfrac{1}{3}|c\rangle \Big\rangle \\
&= \tfrac{1}{3}|a\rangle + \tfrac{2}{9}|b\rangle + \tfrac{4}{9}|c\rangle.
\end{aligned}$$

Exercises

2.3.1 Recall the data / multisets about child ages and blood types in the beginning of Subsection 1.4.1. Compute the associated (empirical) distributions.

Plot these distributions as a graph. How do they compare to the plots (1.16) and (1.17)?

2.3.2 Check that frequentist learning from a constant multiset yields a uniform distribution. And also that frequentist learning is invariant under (non-zero) scalar multiplication for multisets: $\mathit{Flrn}(s \cdot \varphi) = \mathit{Flrn}(\varphi)$ for $s \in \mathbb{R}_{>0}$.

2.3.3 1 Prove that for multisets $\varphi, \psi \in \mathcal{M}_*(X)$ one has:

$$\mathit{Flrn}(\varphi + \psi) = \frac{\|\varphi\|}{\|\varphi\| + \|\psi\|} \cdot \mathit{Flrn}(\varphi) + \frac{\|\psi\|}{\|\varphi\| + \|\psi\|} \cdot \mathit{Flrn}(\psi).$$

This means that when one has already learned $\mathit{Flrn}(\varphi)$ and new data ψ arrives, all probabilities have to be adjusted, as in the above convex sum of distributions.

2 Check that the following formulation for natural multisets of fixed sizes $K > 0$, $L > 0$ is a special case of the previous item.

$$\begin{array}{ccc} \mathcal{N}[K](X) \times \mathcal{N}[L](X) & \xrightarrow{+} & \mathcal{N}[K+L](X) \\ {\scriptstyle \mathit{Flrn} \times \mathit{Flrn}}\downarrow & & \downarrow{\scriptstyle \mathit{Flrn}} \\ \mathcal{D}(X) \times \mathcal{D}(X) & \xrightarrow{\frac{K}{K+L}(-) + \frac{L}{K+L}(-)} & \mathcal{D}(X) \end{array}$$

2.3.4 Show that Diagram (1.30) can be refined to:

$$\begin{array}{ccccc} \mathcal{L}_*(X) & & \xrightarrow{\mathit{supp}} & & \mathcal{P}_{\mathit{fin}}(X) \\ & {\scriptstyle \mathit{acc}}\searrow & & \nearrow{\scriptstyle \mathit{supp}} & \\ & \mathcal{M}_*(X) & \xrightarrow[\mathit{Flrn}]{} & \mathcal{D}(X) & \end{array} \tag{2.16}$$

where $\mathcal{L}_*(X) \subseteq \mathcal{L}(X)$ is the subset of non-empty lists.

2.3.5 Let $c \colon X \to \mathcal{D}(Y)$ be a $\mathcal{D}$-channel and $\varphi \in \mathcal{M}(X)$ be a multiset. Because $\mathcal{D}(Y) \subseteq \mathcal{M}(Y)$ we can also consider c as an $\mathcal{M}$-channel, and use $c \gg\!= \varphi$. Prove that:

$$\mathit{Flrn}(c \gg\!= \varphi) = c \gg\!= \mathit{Flrn}(\varphi) = \mathit{Flrn}(c \gg\!= \mathit{Flrn}(\varphi)).$$

2.3.6 Let X be a finite set and $K \in \mathbb{N}$ be an arbitrary number. Show that for $\sigma \in \mathcal{D}(\mathcal{N}[K+1](X))$ and $\varphi \in \mathcal{N}[K](X)$ one has:

$$\frac{(DD \gg\!= \sigma)(\varphi)}{\binom{\varphi}{}} = \sum_{x \in X} \frac{\sigma(\varphi + 1|x\rangle)}{\binom{\varphi + 1|x\rangle}{}}.$$

2.3.7 Recall the uniform distributions $\mathit{unif}_K \in \mathcal{D}(\mathcal{N}[K](X))$ from Exercise 2.2.9, where the set X is finite. Use Lemma 1.7.5 to prove that $\mathit{Flrn} \gg\!= \mathit{unif}_K = \mathit{unif}_X \in \mathcal{D}(X)$.

2.4 Parallel products

In the very first section of the first chapter we have seen Cartesian, parallel products $X \times Y$ of sets X, Y. Here we shall look at parallel products of states, and also at parallel products of channels. These new products will be written as tensors $\otimes$. They express parallel combination. These tensors exist for $\mathcal{P}$, $\mathcal{M}$ and $\mathcal{D}$, but not for lists $\mathcal{L}$. The reason for this absence will be explained below, in Remark 2.4.3.

In this section we start with a brief uniform description of parallel products, for multiple collection types — in the style of the first chapter — but we shall quickly zoom in on the probabilistic case. Products $\otimes$ of distributions have their own dynamics, due to the requirement that probabilities, also over a product, must add up to one. This means that the two components of a 'joint' distribution, over a product space, can be correlated. Indeed, a joint distribution is typically not equal to the product of its marginals: the whole is more than the product of its parts. It also means, as we shall see in Chapter 6, that updating in one part has effect in other parts: the two parts of a joint distribution 'listen' to each other.

Definition 2.4.1. Tensors, also called parallel products, will be defined first for states, for each collection type separately, and then for channels, in a uniform manner.

1 Let X, Y be arbitrary sets.

   a For $U \in \mathcal{P}(X)$ and $V \in \mathcal{P}(Y)$ we define $U \otimes V \in \mathcal{P}(X \times Y)$ as:

$$U \otimes V \coloneqq \{(x, y) \in X \times Y \mid x \in U \text{ and } y \in V\}.$$

   This product $U \otimes V$ is often written simply as a product of sets $U \times V$, but we prefer to have a separate notation for this product of subsets.

   b For $\varphi \in \mathcal{M}(X)$ and $\psi \in \mathcal{M}(Y)$ we get $\varphi \otimes \psi \in \mathcal{M}(X \times Y)$ as:

$$\varphi \otimes \psi \coloneqq \sum_{x \in X, y \in Y} (\varphi(x) \cdot \psi(y)) \big| x, y \big\rangle \quad \text{that is} \quad (\varphi \otimes \psi)(x, y) = \varphi(x) \cdot \psi(y).$$

   c For $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ we use $\omega \otimes \rho$ as in the previous item, for multisets. This is well-defined, with outcome in $\mathcal{D}(X \times Y)$, since the relevant multiplicities add up to one. This tensor also works for infinite support, *i.e.* for $\mathcal{D}_\infty$.

2 Let $c \colon X \nrightarrow Y$ and $d \colon A \nrightarrow B$ be two channels, both of the same type $T \in \{\mathcal{P}, \mathcal{M}, \mathcal{D}\}$. A channel $c \otimes d \colon X \times A \nrightarrow Y \times B$ is defined via:

$$(c \otimes d)(x, a) \coloneqq c(x) \otimes d(a).$$

The right-hand side uses the tensor product of the appropriate type T, as defined in the previous item.

We shall use tensors not only in binary form $\otimes$, but also in n-ary form $\otimes \cdots \otimes$, both for states and for channels.

We see that tensors $\otimes$ involve the products $\times$ of the underlying domains. A simple illustration of a (probabilistic) tensor product is:

$$\begin{aligned}&(\tfrac{1}{6}|u\rangle + \tfrac{1}{3}|v\rangle + \tfrac{1}{2}|w\rangle) \otimes (\tfrac{3}{4}|0\rangle + \tfrac{1}{4}|1\rangle)\\ &= \tfrac{1}{8}|u,0\rangle + \tfrac{1}{24}|u,1\rangle + \tfrac{1}{4}|v,0\rangle + \tfrac{1}{12}|v,1\rangle + \tfrac{3}{8}|w,0\rangle + \tfrac{1}{8}|w,1\rangle.\end{aligned}$$

These tensor products tend to grow really quickly in size, since the number of entries of the two parts have to be multiplied.

Parallel products are well-behaved, as described below.

Lemma 2.4.2. *1 Transformation of parallel states via parallel channels is the parallel product of the individual transformations:*

$$(c \otimes d) \gg\!= (\omega \otimes \rho) \;=\; (c \gg\!= \omega) \otimes (d \gg\!= \rho).$$

2 Parallel products of channels interact nicely with unit and composition:

$$\mathit{unit} \otimes \mathit{unit} \;=\; \mathit{unit} \qquad (c_1 \otimes d_1) \mathbin{\circ\!\!\circ} (c_2 \otimes d_2) \;=\; (c_1 \mathbin{\circ\!\!\circ} c_2) \otimes (d_1 \mathbin{\circ\!\!\circ} d_2).$$

3 The tensor of trivial, deterministic channels is obtained from their product:

$$\langle f \rangle \otimes \langle g \rangle \;=\; \langle f \times g \rangle$$

where $f \times g = \langle f \circ \pi_1, g \circ \pi_2 \rangle$, see Subsection 1.1.1.

4 The image distribution along a product function is a tensor of images:

$$\mathcal{D}(f \times g)(\omega \otimes \rho) \;=\; \mathcal{D}(f)(\omega) \otimes \mathcal{D}(g)(\rho).$$

Proof. We shall do the proofs for $\mathcal{M}$ and for $\mathcal{D}$ of item (1) at once, using Exercise 2.2.3 (1), and leave the remaining cases to the reader.

$$\begin{aligned}((c \otimes d) \gg\!= (\omega \otimes \rho))(y,b) &= \sum_{x \in X,\, a \in A} (c \otimes d)(x,a)(y,b) \cdot (\omega \otimes \rho)(x,a)\\ &= \sum_{x \in X,\, a \in A} (c(x) \otimes d(a))(y,b) \cdot \omega(x) \cdot \rho(a)\\ &= \sum_{x \in X,\, a \in A} c(x)(y) \cdot d(a)(b) \cdot \omega(x) \cdot \rho(a)\\ &= \left(\sum_{x \in X} c(x)(y) \cdot \omega(x)\right) \cdot \left(\sum_{a \in A} d(a)(b) \cdot \rho(a)\right)\\ &= (c \gg\!= \omega)(y) \cdot (d \gg\!= \rho)(b)\\ &= ((c \gg\!= \omega) \otimes (d \gg\!= \rho))(y,b).\end{aligned}$$

❑

As promised we look into why parallel products don't work for lists.

Remark 2.4.3. Suppose we have two list $[a, b, c]$ and $[u, v]$ and we wish to form their parallel product. Then there are many ways to do so. For instance, two obvious choices are:

$$[\langle a, u\rangle, \langle a, v\rangle, \langle b, u\rangle, \langle b, v\rangle, \langle c, u\rangle, \langle c, v\rangle]$$
$$[\langle a, u\rangle, \langle b, u\rangle, \langle c, u\rangle, \langle a, v\rangle, \langle b, v\rangle, \langle c, v\rangle]$$

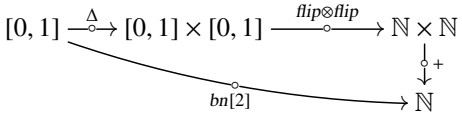
There are many other possibilities. The problem is that there is no canonical choice. Since the order of elements in a list matter, there is no commutativity property which makes all options equivalent, like for multisets. Technically, the tensor for $\mathcal{L}$ does not exist because $\mathcal{L}$ is not a *commutative* (*i.e. monoidal*) monad; this is an early result in category theory going back to [102].

From now on we concentrate on parallel products $\otimes$ for distributions and for probabilistic channels. We illustrate the use of $\otimes$ of probabilistic channels for two standard 'summation' results. We start with a simple property, which is studied in more complicated form in Exercise 2.4.3. Abstractly, it can be understood in terms of *convolutions*. In general, a convolution of parallel maps $f, g\colon X \to Y$ is a composite of the form:

$$X \xrightarrow{\ split\ } X \times X \xrightarrow{\ f\otimes g\ } Y \times Y \xrightarrow{\ join\ } Y \tag{2.17}$$

The *split* and *join* operations depend on the situation. In the next result they are copy and sum.

Lemma 2.4.4. *Recall the flip (or Bernoulli) distribution* $\mathit{flip}(r) = r|1\rangle + (1 - r)|0\rangle$ *and consider it as a channel* $\mathit{flip}\colon [0, 1] \nrightarrow \mathbb{N}$. *The convolution of two such flip's is then a binomial distribution, as described in the following diagram of channels.*



Proof. For $r \in [0, 1]$ we compute:

$$\begin{aligned}
&\mathcal{D}(+)\Big(\mathit{flip}(r) \otimes \mathit{flip}(r)\Big) \\
&= \mathcal{D}(+)\Big(r^2\big|1, 1\big\rangle + r \cdot (1 - r)\big|1, 0\big\rangle + (1 - r) \cdot r)\big|0, 1\big\rangle + (1 - r)^2\big|0, 0\big\rangle\Big) \\
&= r^2\big|2\big\rangle + 2 \cdot r \cdot (1 - r)\big|1\big\rangle + (1 - r)^2\big|0\big\rangle \\
&= \sum_{0\le i\le 2} \binom{2}{i} \cdot r^i \cdot (1 - r)^{2-i}\big|i\big\rangle \;=\; \mathit{bn}[2](r).
\end{aligned}$$

❑

The structure used in this result is worth making explicit. We have introduced probability distributions on *sets*, as sample spaces. It turns out that if this underlying set, say M, happens to be a commutative monoid, then $\mathcal{D}(M)$ also forms a commutative monoid. This makes for instance the set $\mathcal{D}(\mathbb{N})$ of distributions on the natural numbers a commutative monoid — even in two ways, using either the additive or multiplicative monoid structure on $\mathbb{N}$. The construction occurs in [99, p.82], but does not seem to be widely used and/or familiar. It is an instance of an abstract form of convolution in [104, §10]. Since it plays an important role here we make it explicit. It can be described via parallel products $\otimes$ of distributions.

Proposition 2.4.5. *Let $M = (M, +, 0)$ be a commutative monoid. Define a sum and zero element on $\mathcal{D}(M)$ via:*

$$\begin{array}{rcl} \omega + \rho & \coloneqq & \mathcal{D}(+)(\omega \otimes \rho) \\ 0 & \coloneqq & 1|0\rangle \end{array} \qquad \text{using} \qquad \begin{cases} + \colon M \times M \to M \\ 0 \in M \end{cases} \tag{2.18}$$

Alternative descriptions of this sum of distributions are:

$$\omega + \rho = \sum_{a,b \in M} \omega(a) \cdot \rho(b) \big| a + b \big\rangle$$
$$\Big(\sum\nolimits_i r_i \big| a_i \big\rangle\Big) + \Big(\sum\nolimits_j s_j \big| b_j \big\rangle\Big) = \sum\nolimits_{i,j} r_i \cdot s_j \big| a_i + b_j \big\rangle.$$

1 *Via these sums $+, 0$ of distributions, the set $\mathcal{D}(M)$ forms a commutative monoid.*
2 *If $f \colon M \to N$ is a homomorphism of commutative monoids, then so is $\mathcal{D}(f) \colon \mathcal{D}(M) \to \mathcal{D}(N)$.*

Let **CMon** *be the category of commutative monoid, with monoid homomorphisms as arrows between them. The above two items tell that the distribution functor $\mathcal{D} \colon$* **Sets** $\to$ **Sets** *can be restricted to a functor $\mathcal{D} \colon$* **CMon** $\to$ **CMon** *in a commuting diagram:*

$$\begin{array}{ccc} \mathbf{CMon} & \xrightarrow{\ \mathcal{D}\ } & \mathbf{CMon} \\ \downarrow & & \downarrow \\ \mathbf{Sets} & \xrightarrow{\ \mathcal{D}\ } & \mathbf{Sets} \end{array} \tag{2.19}$$

The vertical arrows 'forget' the monoid structure, by sending a monoid to its underlying set.

Exercise 2.4.4 below contains some illustrations of this definition. In Exercise 2.4.12 we shall see that the restricted (or lifted) functor $\mathcal{D} \colon$ **CMon** $\to$ **CMon** is also a monad. The same construction works for $\mathcal{D}_\infty$.

Proof. 1 Commutativity and associativity of + on $\mathcal{D}(M)$ follow from commutativity and associativity of + on M, and of multiplication $\cdot$ on $[0, 1]$. Next,

$$\omega + 0 = \omega + 1|0\rangle = \sum_{a\in M} \omega(a) \cdot 1 \,\big|a + 0\big\rangle = \sum_{a\in M} \omega(a) \,\big|a\big\rangle = \omega.$$

2 Clearly, $\mathcal{D}(f)(1|0\rangle) = 1|f(0)\rangle = 1|0\rangle$, and:

$$\begin{aligned}
\mathcal{D}(f)(\omega + \rho) &= \Big(\mathcal{D}(f) \circ \mathcal{D}(+)\Big)(\omega \otimes \rho) \\
&= \Big(\mathcal{D}(+) \circ \mathcal{D}(f \times f)\Big)(\omega \otimes \rho) \\
&= \mathcal{D}(+)\Big(\mathcal{D}(f)(\omega) \otimes \mathcal{D}(f)(\rho)\Big) \qquad \text{by Lemma 2.4.2 (4)} \\
&= \mathcal{D}(f)(\omega) + \mathcal{D}(f)(\rho). \qquad \blacksquare
\end{aligned}$$

As mentioned, we use this sum of distributions in Lemma 2.4.4. We may now reformulate it as:

$$\mathit{flip}(r) + \mathit{flip}(r) = \mathit{bn}[2](r).$$

We consider a similar property of Poisson distributions $\mathit{pois}[\lambda]$, see (2.7). This property is commonly expressed in terms of random variables X_i as: if $X_1 \sim \mathit{pois}[\lambda_1]$ and $X_2 \sim \mathit{pois}[\lambda_2]$ then $X_1 + X_2 \sim \mathit{pois}[\lambda_1 + \lambda_2]$. We have not discussed random variables yet, nor the meaning of $\sim$, but we do not need them for the channel-based reformulation below.

Recall that the Poisson distribution has infinite support, so that we need to use $\mathcal{D}_\infty$ instead of $\mathcal{D}$, see Remark 2.1.4, but that difference is immaterial here. We now use the mapping $\lambda \mapsto \mathit{pois}[\lambda]$ as a function $\mathbb{R}_{\geq 0} \to \mathcal{D}_\infty(\mathbb{N})$ and as a $\mathcal{D}_\infty$-channel $\mathit{pois} \colon \mathbb{R}_{\geq 0} \nrightarrow \mathbb{N}$.

Proposition 2.4.6. *The Poisson channel* $\mathit{pois} \colon \mathbb{R}_{\geq 0} \nrightarrow \mathbb{N}$ *is a homomorphism of monoids.*

Thus, for $\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$,

$$\mathit{pois}[\lambda_1 + \lambda_2] = \mathit{pois}[\lambda_1] + \mathit{pois}[\lambda_2] \qquad \text{and} \qquad \mathit{pois}[0] = 1|0\rangle.$$

We can also express this via commutation of the following diagrams of channels.

$$\begin{array}{ccc}
\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} & \overset{+}{\longrightarrow} & \mathbb{R}_{\geq 0} \\
{\scriptstyle \mathit{pois} \otimes \mathit{pois}}\downarrow & & \downarrow{\scriptstyle \mathit{pois}} \\
\mathbb{N} \times \mathbb{N} & \overset{+}{\longrightarrow} & \mathbb{N}
\end{array}
\qquad\qquad
\begin{array}{ccc}
1 & \overset{0}{\longrightarrow} & \mathbb{R}_{\geq 0} \\
{\scriptstyle \mathit{unit}}\| & & \downarrow{\scriptstyle \mathit{pois}} \\
1 & \overset{0}{\longrightarrow} & \mathbb{N}
\end{array}$$

Proof. We first do preservation of sums +, for which we pick arbitrary $\lambda_1, \lambda_2 \in$

$\mathbb{R}_{\geq 0}$ and $k \in \mathbb{N}$.

$$\begin{aligned}
(pois[\lambda_1] + pois[\lambda_2])(k) &\overset{(2.18)}{=} \mathcal{D}(+)(pois[\lambda_1] \otimes pois[\lambda_2])(k) \\
&= \sum_{k_1,k_2,k_1+k_2=k} (pois[\lambda_1] \otimes pois[\lambda_2])(k_1, k_2) \\
&= \sum_{0\leq m\leq k} pois[\lambda_1](m) \cdot pois[\lambda_2](k-m) \\
&\overset{(2.7)}{=} \sum_{0\leq m\leq k} \left(e^{-\lambda_1} \cdot \frac{\lambda_1^m}{m!}\right) \cdot \left(e^{-\lambda_2} \cdot \frac{\lambda_2^{k-m}}{(k-m)!}\right) \\
&= \sum_{0\leq m\leq k} \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot \frac{k!}{m!(k-m)!} \cdot \lambda_1^m \cdot \lambda_2^{k-m} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot \sum_{0\leq m\leq k} \binom{k}{m} \cdot \lambda_1^m \cdot \lambda_2^{k-m} \\
&\overset{(1.26)}{=} \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot (\lambda_1 + \lambda_2)^k \\
&\overset{(2.7)}{=} pois[\lambda_1 + \lambda_2](k).
\end{aligned}$$

Finally, in the expression $pois[0] = \sum_k e^0 \cdot \frac{0^k}{k!}|k\rangle$ everything vanishes except for $k = 0$, since only $0^0 = 1$. Hence $pois[0] = 1|0\rangle$. ❑

The next illustration shows that tensors for different collection types are related.

Proposition 2.4.7. *The maps* $supp\colon \mathcal{D} \to \mathcal{P}$ *and* $Flrn\colon \mathcal{M}_* \to \mathcal{D}$ *commute with tensors, in the sense that:*

$$supp(\sigma \otimes \tau) = supp(\sigma) \otimes supp(\tau) \qquad Flrn(\varphi \otimes \psi) = Flrn(\varphi) \otimes Flrn(\psi).$$

This says that support and frequentist learning are 'monoidal' natural transformations.

Proof. We shall do the *Flrn*-case and leave *supp* as exercise. We first note that:

$$\begin{aligned}
\big\|\varphi \otimes \psi\big\| = \textstyle\sum_z (\varphi \otimes \psi)(z) &= \textstyle\sum_{x,y} (\varphi \otimes \psi)(x, y) \\
&= \textstyle\sum_{x,y} \varphi(x) \cdot \psi(y) \\
&= (\textstyle\sum_x \varphi(x)) \cdot (\textstyle\sum_y \psi(y)) = \big\|\varphi\big\| \cdot \big\|\psi\big\|.
\end{aligned}$$

Then:

$$\begin{aligned}
Flrn(\varphi \otimes \psi)(x, y) = \frac{(\varphi \otimes \psi)(x, y)}{\|\varphi \otimes \psi\|} &= \frac{\varphi(x)}{\|\varphi\|} \cdot \frac{\psi(y)}{\|\psi\|} \\
&= Flrn(\varphi)(x) \cdot Flrn(\psi)(y) \\
&= (Flrn(\varphi) \otimes Flrn(\psi))(x, y).
\end{aligned}$$ ❑

We can form a K-ary product X^K for a set X. But also for a distribution $\omega \in \mathcal{D}(X)$ we can form $\omega^K = \omega \otimes \cdots \otimes \omega \in \mathcal{D}(X^K)$. This lies at the heart of the notion of 'independent and identical distributions'. We define a separate function for this construction:

$$\mathcal{D}(X) \xrightarrow{iid[K]} \mathcal{D}(X^K) \quad \text{with} \quad iid[K](\omega) = \underbrace{\omega \otimes \cdots \otimes \omega}_{K \text{ times}}. \tag{2.20}$$

We can describe $iid[K]$ diagrammatically as composite, making the copying of states explicit:

$$\begin{array}{ccc} \mathcal{D}(X) & \xrightarrow{\quad iid[K] \quad} & \mathcal{D}(X^K) \\ & {\scriptstyle \Delta} \searrow \quad \nearrow {\scriptstyle \bigotimes[K]} & \\ & \mathcal{D}(X)^K & \end{array} \quad \text{where} \quad \begin{array}{l} \bigotimes[K](\omega_1, \ldots, \omega_K) \\ := \omega_1 \otimes \cdots \otimes \omega_K. \end{array} \tag{2.21}$$

We often omit the parameter K when it is clear from the context. This map iid will pop-up occasionally in the sequel. At this stage we only mention a few of its properties, in combination with zipping.

Lemma 2.4.8. *Consider the* iid *maps from* (2.20)*, and also the* zip *function from Exercise 1.1.7.*

1 *These* iid *maps are natural, which we shall describe in two ways. For a function $f\colon X \to Y$ and for a channel $c\colon X \nrightarrow Y$ the following two diagrams commutes.*

$$\begin{array}{ccc} \mathcal{D}(X) & \xrightarrow{iid} & \mathcal{D}(X^K) \\ {\scriptstyle \mathcal{D}(f)}\downarrow & & \downarrow{\scriptstyle \mathcal{D}(f^K)} \\ \mathcal{D}(Y) & \xrightarrow{iid} & \mathcal{D}(Y^K) \end{array} \qquad \begin{array}{ccc} \mathcal{D}(X) & \xrightarrow{iid} & X^K \\ {\scriptstyle c \gg= (-)}\downarrow & & \downarrow{\scriptstyle c^K = c \otimes \cdots \otimes c} \\ \mathcal{D}(Y) & \xrightarrow{iid} & Y^K \end{array}$$

2 *The* iid *maps interact suitably with tensors of distributions, as expressed by the following diagram.*

$$\begin{array}{ccc} \mathcal{D}(X) \times \mathcal{D}(Y) & \xrightarrow{iid \otimes iid} & X^K \times Y^K \\ {\scriptstyle \otimes}\downarrow & & \downarrow{\scriptstyle \langle zip \rangle} \\ \mathcal{D}(X \times Y) & \xrightarrow{iid} & (X \times Y)^K \end{array}$$

3 *The* zip *function commutes with tensors in the following way.*

$$\begin{array}{ccccc} \mathcal{D}(X)^K \times \mathcal{D}(Y)^K & \xrightarrow{\bigotimes \times \bigotimes} & \mathcal{D}(X^K) \times \mathcal{D}(Y^K) & \xrightarrow{\otimes} & \mathcal{D}(X^K \times Y^K) \\ {\scriptstyle zip}\downarrow & & & & \downarrow{\scriptstyle \mathcal{D}(zip)} \\ (\mathcal{D}(X) \times \mathcal{D}(Y))^K & \xrightarrow{\otimes^K} & \mathcal{D}(X \times Y)^K & \xrightarrow{\bigotimes} & \mathcal{D}((X \times Y)^K) \end{array}$$

Proof. 1 Commutation of the diagram on the left, for a function f, is a direct consequence of Lemma 2.4.2 (4), so we concentrate on the one on the right, for a channel $c\colon X \nrightarrow Y$. We use Lemma 2.4.2 (1) in:

$$\begin{aligned}(c^K \circ \mathit{iid})(\omega) &= (c \otimes \cdots \otimes c) \gg\!= (\omega \otimes \cdots \otimes \omega) \\ &= (c \gg\!= \omega) \otimes \cdots \otimes (c \gg\!= \omega) \\ &= \mathit{iid}(c \gg\!= \omega) \\ &= (\mathit{iid} \circ (c \gg\!= (-)))(\omega).\end{aligned}$$

2 For $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$,

$$\begin{aligned}(\langle \mathit{zip} \rangle \circ (\mathit{iid} \otimes \mathit{iid}))(\omega, \rho) &= \mathcal{D}(\mathit{zip})(\omega^K \otimes \rho^K) \\ &= \sum_{\vec{x} \in X^K} \sum_{\vec{y} \in Y^K} \omega^K(\vec{x}) \cdot \rho^K(\vec{y}) \big| \mathit{zip}(\vec{x}, \vec{z}) \big\rangle \\ &= \sum_{\vec{z} \in (X \times Y)^K} (\omega \otimes \rho)^K(\vec{z}) \big| \vec{z} \big\rangle \\ &= (\omega \otimes \rho)^K \\ &= \mathit{iid}(\omega \otimes \rho).\end{aligned}$$

3 For distributions $\omega_i \in \mathcal{D}(X)$, $\rho_i \in \mathcal{D}(Y)$ and elements $x_i \in X$, $y_i \in Y$ we elaborate:

$$\begin{aligned}&(\mathcal{D}(\mathit{zip}) \circ \otimes \circ (\textstyle\bigotimes \times \bigotimes))([\omega_1, \ldots, \omega_K], [\rho_1, \ldots, \rho_K]) \\ &\qquad\qquad ([(x_1, y_1), \ldots, (x_K, y_K)]) \\ &= \mathcal{D}(\mathit{zip})\Big((\omega_1 \otimes \cdots \otimes \omega_K) \otimes (\rho_1 \otimes \cdots \rho_K)\Big)([(x_1, y_1), \ldots, (x_K, y_K)]) \\ &= (\omega_1 \otimes \cdots \otimes \omega_K)([x_1, \ldots, x_K]) \cdot (\rho_1 \otimes \cdots \otimes \rho_K)([y_1, \ldots, y_K]) \\ &= \omega_1(x_1) \cdot \ldots \cdot \omega_K(x_K) \cdot \rho_1(y_1) \cdot \ldots \cdot \rho_K(y_K) \\ &= \omega_1(x_1) \cdot \rho_1(y_1) \cdot \ldots \cdot \omega_K(x_K) \cdot \rho_K(y_K) \\ &= (\omega_1 \otimes \rho_1)(x_1, y_1) \cdot \ldots \cdot (\omega_K \otimes \rho_K)(x_K, y_K) \\ &= \Big((\omega_1 \otimes \rho_1) \otimes \cdots \otimes (\omega_K \otimes \rho_K)\Big)([(x_1, y_1), \ldots, (x_K, y_K)]) \\ &= (\textstyle\bigotimes \circ \otimes^K)([(\omega_1, \rho_1), \ldots, (\omega_K, \rho_K)])([(x_1, y_1), \ldots, (x_K, y_K)]) \\ &= (\textstyle\bigotimes \circ \otimes^K \circ \mathit{zip})([\omega_1, \ldots, \omega_K], [\rho_1, \ldots, \rho_K]) \\ &\qquad\qquad ([(x_1, y_1), \ldots, (x_K, y_K)]).\end{aligned}$$

❑

Exercises

2.4.1 Prove the equation for *supp* in Proposition 2.4.7.

2.4.2 Show that tensoring of multisets is linear, in the sense that for $\sigma \in \mathcal{M}(X)$ the operation 'tensor with σ' $\sigma \otimes (-)\colon \mathcal{M}(Y) \to \mathcal{M}(X \times Y)$ is

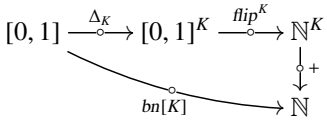
linear w.r.t. the cone structure of Lemma 1.4.2 (2): for $\tau_1, \dots, \tau_n \in \mathcal{D}(Y)$ and $r_1, \dots, r_n \in \mathbb{R}_{\geq 0}$ one has:

$$\sigma \otimes \Big(r_1 \cdot \tau_1 + \cdots + r_n \cdot \tau_n \Big) = r_1 \cdot (\sigma \otimes \tau_1) + \cdots + r_n \cdot (\sigma \otimes \tau_n).$$

The same hold in the other coordinate, for $(-)\otimes\tau$. As a special case we obtain that when σ is a probability distribution, then $\sigma\otimes(-)$ preserves convex sums of distributions.

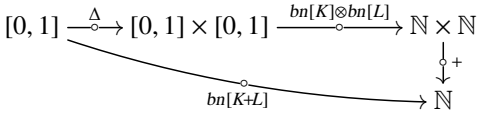
2.4.3 Extend Lemma 2.4.4 in the following two ways.

1 Show that the K-fold convolution (2.17) of *flip*'s is a binomial of size K, as in:



where Δ_K is a K-fold copy and $+$ returns the sum of a K-tuple of numbers.

2 Show that binomials are closed under convolution, in the following sense: for numbers $K, L \in \mathbb{N}$,



Hint: Remember Vandermonde's binary formula (1.29).

2.4.4 The set of natural numbers $\mathbb{N}$ has two commutative monoid structures, one additive with $+, 0$, and one multiplicative with $\cdot, 1$. Accordingly Proposition 2.4.5 gives two commutative monoid structures on $\mathcal{D}(\mathbb{N})$, namely:

$$\omega + \rho = \mathcal{D}(+)(\omega \otimes \rho) \qquad \text{and} \qquad \omega \star \rho = \mathcal{D}(\cdot)(\omega \otimes \rho).$$

Consider the following three distributions on $\mathbb{N}$.

$$\rho_1 = \tfrac{1}{2}|0\rangle + \tfrac{1}{3}|1\rangle + \tfrac{1}{6}|2\rangle \quad \rho_2 = \tfrac{1}{2}|0\rangle + \tfrac{1}{2}|1\rangle \quad \omega = \tfrac{2}{3}|0\rangle + \tfrac{1}{3}|1\rangle.$$

Show consecutively:

1 $\rho_1 + \rho_2 = \frac{1}{4}|0\rangle + \frac{5}{12}|1\rangle + \frac{1}{4}|2\rangle + \frac{1}{12}|3\rangle$;
2 $\omega \star (\rho_1 + \rho_2) = \frac{3}{4}|0\rangle + \frac{5}{36}|1\rangle + \frac{1}{12}|2\rangle + \frac{1}{36}|3\rangle$;
3 $\omega \star \rho_1 = \frac{5}{6}|0\rangle + \frac{1}{9}|1\rangle + \frac{1}{18}|2\rangle$;
4 $\omega \star \rho_2 = \frac{5}{6}|0\rangle + \frac{1}{6}|1\rangle$;
5 $(\omega \star \rho_1) + (\omega \star \rho_2) = \frac{25}{36}|0\rangle + \frac{25}{108}|1\rangle + \frac{7}{108}|2\rangle + \frac{1}{108}|3\rangle$.

Observe that $\star$ does not distribute over $+$ on $\mathcal{D}(\mathbb{N})$. More generally, conclude that the construction of Proposition 2.4.5 does not extend to commutative semirings.

2.4.5 Recall that for $N \in \mathbb{N}_{>0}$ we write $N = \{0, 1, \ldots, N-1\}$ for the set of natural numbers (strictly) below N. It is an additive monoid, via addition modulo N. As such it is sometimes written as $\mathbb{Z}_N$ or as $\mathbb{Z}/N\mathbb{Z}$. Prove that:

$$\mathit{unif}_N + \mathit{unif}_N = \mathit{unif}_N, \qquad \text{with } + \text{ from Proposition 2.4.5.}$$

You may wish to check this equation first for $N = 4$ or $N = 5$. It works for the modular sum, not for the ordinary sum (on $\mathbb{N}$), as one can see from the sum *dice* + *dice*, when *dice* is considered as distribution on $\mathbb{N}$, see also Example 4.5. See [148] for more info.

2.4.6 Let $M = (M, +, 0)$ and $N = (N, +, 0)$ be two commutative monoids, so that $\mathcal{D}(M)$ and $\mathcal{D}(N)$ are also commutative monoids. Let channel $c \colon M \to \mathcal{D}(N)$ be a homomorphism of monoids. Show that state transformation $c \gg\!= (-) \colon \mathcal{D}(M) \to \mathcal{D}(N)$ is then also a homomorphism of monoids.

2.4.7 For sets X, Y with arbitrary elements $a \in X$, $b \in Y$ and with distributions $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$ we define *strength* maps $\mathit{st}_1 \colon \mathcal{D}(X) \times Y \to \mathcal{D}(X \times Y)$ and $\mathit{st}_2 \colon X \times \mathcal{D}(Y) \to \mathcal{D}(X \times Y)$ via:

$$\begin{aligned} \mathit{st}_1(\sigma, b) &\coloneqq \sigma \otimes \mathit{unit}(b) = \sum_{x \in X} \sigma(x) \,\big|x, b\big\rangle \\ \mathit{st}_2(a, \tau) &\coloneqq \mathit{unit}(a) \otimes \tau = \sum_{y \in Y} \tau(y) \,\big|a, y\big\rangle. \end{aligned} \tag{2.22}$$

Show that the tensor $\sigma \otimes \tau$ can be reconstructed from these two strength maps via the following diagram:

$$\begin{array}{ccccc} \mathcal{D}(X \times \mathcal{D}(Y)) & \xleftarrow{\mathit{st}_1} & \mathcal{D}(X) \times \mathcal{D}(Y) & \xrightarrow{\mathit{st}_2} & \mathcal{D}(\mathcal{D}(X) \times Y) \\ {\scriptstyle \mathcal{D}(\mathit{st}_2)}\downarrow & & \otimes\downarrow & & \downarrow{\scriptstyle \mathcal{D}(\mathit{st}_1)} \\ \mathcal{D}(\mathcal{D}(X \times Y)) & \xrightarrow{\mathit{flat}} & \mathcal{D}(X \times Y) & \xleftarrow{\mathit{flat}} & \mathcal{D}(\mathcal{D}(X \times Y)) \end{array}$$

2.4.8 Consider a function $f \colon X \to M$ where X is an ordinary set and M is a monoid. We can *add noise* to f via a channel $c \colon X \nrightarrow M$. The result is a channel $\mathit{noise}(f, c) \colon X \nrightarrow M$ given by:

$$\mathit{noise}(f, c) \coloneqq \langle f \rangle + c.$$

This formula promotes f to the deterministic channel $\langle f \rangle = \mathit{unit} \circ f$ and uses the sum $+$ of distributions from (2.18) in a pointwise manner.

1 Check that we can concretely describe this noise channel as:

$$noise(f, c)(x) = \sum_{y \in Y} c(x)(y) \,\big|f(x) + y\big\rangle.$$

2 Show also that it can be described abstractly via strength (from the previous exercise) as composite:

$$noise(f, c) = \Big(X \xrightarrow{\langle f, c\rangle} M \times \mathcal{D}(M) \xrightarrow{st_2} \mathcal{D}(M \times M) \xrightarrow{\mathcal{D}(+)} \mathcal{D}(M)\Big).$$

2.4.9 Show that the following diagram commutes.

$$\begin{array}{ccc} \mathcal{D}(\mathcal{D}(X)) & \xrightarrow{iid} & \mathcal{D}(X)^K \\ {\scriptstyle flat}\downarrow & & \downarrow{\scriptstyle \bigotimes} \\ \mathcal{D}(X) & \xrightarrow{iid} & X^K \end{array} \qquad \text{that is} \qquad \bigotimes \gg\!= iid(\Omega) = iid(flat(\Omega)).$$

2.4.10 Show that the big tensor $\bigotimes\colon \mathcal{D}(X)^K \to \mathcal{D}(X^K)$ from (2.21) commutes with unit and flatten, as described below.

$$\begin{array}{ccc} X^K & & \\ {\scriptstyle unit^K}\downarrow & \searrow^{unit} & \\ \mathcal{D}(X)^K & \xrightarrow{\bigotimes} & \mathcal{D}(X^K) \end{array} \qquad \begin{array}{ccccc} \mathcal{D}^2(X)^K & \xrightarrow{\bigotimes} & \mathcal{D}(\mathcal{D}(X)^K) & \xrightarrow{\mathcal{D}(\bigotimes)} & \mathcal{D}^2(X^K) \\ {\scriptstyle flat^K}\downarrow & & & & \downarrow{\scriptstyle flat} \\ \mathcal{D}(X)^K & & \xrightarrow{\qquad\qquad\bigotimes\qquad\qquad} & & \mathcal{D}(X^K) \end{array}$$

Abstractly this shows that the functor $(-)^K$ distributes over the monad $\mathcal{D}$.

2.4.11 Check that Lemma 2.4.2 (4) can be read as: the tensor $\otimes$ of distributions, considered as a function $\otimes\colon \mathcal{D}(X) \times \mathcal{D}(Y) \to \mathcal{D}(X \times Y)$, is a natural transformation from the composite functor:

$$\mathbf{Sets} \times \mathbf{Sets} \xrightarrow{\mathcal{D} \times \mathcal{D}} \mathbf{Sets} \times \mathbf{Sets} \xrightarrow{\times} \mathbf{Sets}$$

to the functor

$$\mathbf{Sets} \times \mathbf{Sets} \xrightarrow{\times} \mathbf{Sets} \xrightarrow{\mathcal{D}} \mathbf{Sets}.$$

2.4.12 Consider the situation described in Proposition 2.4.5, with a commutative monoid M, and induced monoid structure on $\mathcal{D}(M)$.

1 Check that $1|a\rangle + 1|b\rangle = 1|a + b\rangle$, for all $a, b \in M$. This says that $unit\colon M \to \mathcal{D}(M)$ is a homomorphism of monoids, which can also be expressed via commutation of the diagram:

$$\begin{array}{ccc} M \times M & \xrightarrow{unit \times unit} & \mathcal{D}(M) \times \mathcal{D}(M) \\ {\scriptstyle +}\downarrow & & \downarrow{\scriptstyle +} \\ M & \xrightarrow{\qquad unit \qquad} & \mathcal{D}(M) \end{array}$$

2 Check also that $\mathit{flat} \colon \mathcal{D}(\mathcal{D}(M)) \to \mathcal{D}(M)$ is a monoid homomorphism. This means that for $\Omega, \Theta \in \mathcal{D}(\mathcal{D}(M))$ one has:

$$\mathit{flat}(\Omega) + \mathit{flat}(\Theta) = \mathit{flat}(\Omega + \Theta) \quad \text{and} \quad \mathit{flat}(1\big|1|0\rangle\rangle) = 1|0\rangle.$$

The sum + on the on the left-hand side is the one in $\mathcal{D}(M)$, from the beginning of this exercise. The sum + on the right-hand side is the one in $\mathcal{D}(\mathcal{D}(M))$, using that $\mathcal{D}(M)$ is a commutative monoid — and thus $\mathcal{D}(\mathcal{D}(M))$ too.

3 Check the functor $\mathcal{D} \colon \mathbf{CMon} \to \mathbf{CMon}$ in (2.19) is also a monad.

2.5 Projecting and copying

For cartesian products $\times$ there are two projection functions $\pi_1 \colon X_1 \times X_2 \to X_1$, $\pi_2 \colon X_1 \times X_2 \to X_2$ and a diagonal function $\Delta \colon X \to X \times X$ for copying, see Section 1.1. There are analogues for tensors $\otimes$ of distributions. But they behave differently. Since these differences are fundamental in probability theory we devote a separate section to them.

2.5.1 Marginalisation and entwinedness

In Section 1.1 we have seen projection maps $\pi_i \colon X_1 \times \cdots \times X_n \to X_i$ for Cartesian products of sets. Since these π_i are ordinary functions, they can be turned into channels $\langle \pi_i \rangle = \mathit{unit} \circ \pi_i \colon X_1 \times \cdots \times X_n \to\!\!\!\!\bullet\; X_i$. We will be sloppy and typically omit these brackets $\langle - \rangle$. The kind of arrow, $\to$ or $\to\!\!\!\!\bullet$, or the type of operation at hand, will then indicate whether π_i is used as function or as channel. State transformation $\pi_i \gg\!= (-) = \mathcal{D}(\pi_i) \colon \mathcal{D}(X_1 \times \cdots \times X_n) \to \mathcal{D}(X_i)$ with such projections is called *marginalisation*. It is given by summing over all variables, except the one that we wish to keep: for $y \in X_i$ we get, by Definition 2.1.2,

$$\begin{aligned} &(\pi_i \gg\!= \omega)(y) \\ &= \sum_{x_1 \in X_1, \ldots, x_{i-1} \in X_{i-1}} \; \sum_{x_{i+1} \in X_{i+1}, \ldots, x_n \in X_n} \omega(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n). \end{aligned} \tag{2.23}$$

Below we shall introduce special notation for such marginalisation, but first we look at some of its properties.

The next result only works in the probabilistic case, for $\mathcal{D}$. The exercises below will provide counterexamples for $\mathcal{P}$ and $\mathcal{M}$.

Lemma 2.5.1. *1 Projection channels take parallel products of probabilistic states apart, that is, for $\omega_i \in \mathcal{D}(X_i)$ we have:*

$$\pi_i \gg\!= (\omega_1 \otimes \cdots \otimes \omega_n) \;=\; \mathcal{D}(\pi_i)(\omega_1 \otimes \cdots \otimes \omega_n) \;=\; \omega_i.$$

Thus, marginalisation of parallel product yields its components.

2 Similarly, projection channels commute with parallel products of probabilistic channels, in the following manner:

$$\begin{array}{ccc} X_1 \times \cdots \times X_n & \xrightarrow{\;c_1 \otimes \cdots \otimes c_n\;} & Y_1 \times \cdots \times Y_n \\ {\scriptstyle \pi_i}\downarrow & & \downarrow{\scriptstyle \pi_i} \\ X_i & \xrightarrow[\;c_i\;]{} & Y_i \end{array}$$

Proof. We only do item (1), since item (2) then follows easily, using that parallel products of channels are defined pointwise, see Definition 2.4.1 (2). The first equation in item (1) follows from Lemma 1.8.3 (4), which yields $\langle \pi_i \rangle \gg\!= (-) = \mathcal{D}(\pi_i)(-)$. We restrict to the special case where $n = 2$ and $i = 1$. Then:

$$\begin{array}{rcl} \mathcal{D}(\pi_1)(\omega_1 \otimes \omega_2)(x) & \overset{(2.23)}{=} & \sum_y (\omega_1 \otimes \omega_2)(x, y) \\ & = & \sum_y \omega_1(x) \cdot \omega_2(y) \\ & = & \omega_1(x) \cdot \big(\sum_y \omega_2(y)\big) \;=\; \omega_1(x) \cdot 1 \;=\; \omega_1(x). \end{array}$$ ❑

The last line of this proof relies on the fact that probabilistic states (distributions) involve a convex sum, with multiplicities adding up to one. This does not work for subsets or multisets, see Exercise 2.5.1 below.

We introduce special, post-fix notation for marginalisation via 'masks'. It corresponds to the idea of listing only the relevant variables in traditional notation, where a distribution on a product set is often written as $\omega(x, y)$ and its first marginal as $\omega(x)$.

Definition 2.5.2. A *mask* M is a finite list of 0's and 1's, that is an element $M \in \mathcal{L}(\{0, 1\})$. For a state ω of type $T \in \{\mathcal{P}, \mathcal{M}, \mathcal{D}\}$ on $X_1 \times \cdots \times X_n$ and a mask M of length n we write:

$$\omega M$$

for the marginal with mask M. Informally, it keeps all the parts from ω at a position where there is 1 in M and it projects away parts where there is 0. This is best illustrated via an example:

$$\begin{array}{rcl} \omega[1, 0, 1, 0, 1] & = & T(\langle \pi_1, \pi_3, \pi_5 \rangle)(\omega) \;\in\; T(X_1 \times X_3 \times X_5) \\ & = & \langle \pi_1, \pi_3, \pi_5 \rangle \gg\!= \omega. \end{array}$$

More generally, for a channel $c\colon Y \nrightarrow X_1 \times \cdots \times X_n$ and a mask M of length n we use pointwise marginalisation via the same postcomposition:

$$cM \qquad \text{is the channel} \qquad y \longmapsto c(y)M.$$

With Cartesian products one can take an arbitrary tuple $t \in X \times Y$ apart into $\pi_1(t) \in X$, $\pi_2(t) \in Y$. By re-assembling these parts the original tuple is recovered: $\langle \pi_1(t), \pi_2(t)\rangle = t$. This does not work for collections: a joint state is typically not the parallel product of its marginals, see Example 2.5.4 below. We introduce a special name for this.

Definition 2.5.3. A joint state ω on $X \times Y$ is called *non-entwined* if it is the parallel product of its marginals:

$$\omega = \omega[1,0] \otimes \omega[0,1].$$

Otherwise it is called *entwined*. This notion of entwinedness may be formulated with respect to n-ary states, via a mask, see for example Exercise 2.5.2, but it may then need some re-arrangement of components.

Lemma 2.5.1 (1) shows that a probabilistic product state of the form $\omega_1 \otimes \omega_2$ is non-entwined. But in general, joint states are entwined, so that the different parts are correlated and can influence each other. This is a mechanism that will play an important role in the sequel. A joint distribution is more than the product of its parts, and its different parts can influence each other.

Non-entwined states are called separable in [25]. Sometimes they are also called independent, although independence is also used for random variables. We like to have separate terminology for states only, so we use the phrase (non-)entwinedness, which is a new expression. Independence for random variables is described in Section 5.4.

Example 2.5.4. Take $X = \{u, v\}$ and $A = \{a, b\}$ and consider the state $\omega \in \mathcal{D}(X \times A)$ given by:

$$\omega = \tfrac{1}{8}|u,a\rangle + \tfrac{1}{4}|u,b\rangle + \tfrac{3}{8}|v,a\rangle + \tfrac{1}{4}|v,b\rangle.$$

We claim that ω is entwined. Indeed, ω has first and second marginals $\omega[1,0] = \mathcal{D}(\pi_1)(\omega) \in \mathcal{D}(X)$ and $\omega[0,1] = \mathcal{D}(\pi_2)(\omega) \in \mathcal{D}(A)$, namely:

$$\omega[1,0] = \tfrac{3}{8}|u\rangle + \tfrac{5}{8}|v\rangle \qquad \text{and} \qquad \omega[0,1] = \tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle.$$

The original state ω differs from the product of its marginals:

$$\omega[1,0] \otimes \omega[0,1] = \tfrac{3}{16}|u,a\rangle + \tfrac{3}{16}|u,b\rangle + \tfrac{5}{16}|v,a\rangle + \tfrac{5}{16}|v,b\rangle.$$

This entwinedness follows from a general characterisation, see Exercise 2.5.9 below.

2.5.2 Copying

For an arbitrary set X there is a K-ary *copy* function $\Delta[K] \colon X \to X^K = X \times \cdots \times X$ (K times), given as:

$$\Delta[K](x) \coloneqq \langle x, \ldots, x \rangle \quad (K \text{ times } x).$$

We often omit the subscript K, when it is clear from the context, especially when $K = 2$. This copy function can be turned into a copy channel $\langle \Delta[K] \rangle = \mathit{unit} \circ \Delta_K \colon X \nrightarrow X^K$, but, recall, we often omit writing $\langle - \rangle$ for simplicity. These $\Delta[K]$ are alternatively called *copiers* or *diagonals*.

As functions one has $\pi_i \circ \Delta[K] = \mathit{id}$, and thus as channels $\pi_i \circledcirc \Delta[K] = \mathit{unit}$.

Fact 2.5.5. State transformation with a copier $\Delta \gg\!= \omega$ differs from the tensor product $\omega \otimes \omega$. In general, for $K \geq 2$,

$$\Delta[K] \gg\!= \omega \;\neq\; \underbrace{\omega \otimes \cdots \otimes \omega}_{K \text{ times}} \overset{(2.20)}{=} \mathit{iid}[K](\omega).$$

The following simple example illustrates this fact. For clarity we now write the brackets $\langle - \rangle$ explicitly. First,

$$\langle \Delta \rangle \gg\!= \left(\tfrac{1}{3}|0\rangle + \tfrac{2}{3}|1\rangle\right) = \tfrac{1}{3}\big|\Delta(0)\big\rangle + \tfrac{2}{3}\big|\Delta(1)\big\rangle = \tfrac{1}{3}\big|0,0\big\rangle + \tfrac{2}{3}\big|1,1\big\rangle. \tag{2.24}$$

In contrast:

$$\begin{aligned} &\left(\tfrac{1}{3}|0\rangle + \tfrac{2}{3}|1\rangle\right) \otimes \left(\tfrac{1}{3}|0\rangle + \tfrac{2}{3}|1\rangle\right) \\ &= \tfrac{1}{9}|0,0\rangle + \tfrac{2}{9}|0,1\rangle + \tfrac{2}{9}|1,0\rangle + \tfrac{4}{9}|1,1\rangle. \end{aligned} \tag{2.25}$$

One might expect a commuting diagram like in Lemma 2.5.1 (2) for copiers, but that does not work: diagonals do not commute with arbitrary channels, see Exercise 2.5.10 below for a counterexample. Copiers do commute with deterministic channels of the form $\langle f \rangle = \mathit{unit} \circ f$, as in:

$$\Delta \circledcirc \langle f \rangle = (\langle f \rangle \otimes \langle f \rangle) \circledcirc \Delta \qquad \text{because} \qquad \Delta \circ f = (f \times f) \circ \Delta.$$

In fact, this commutation with diagonals may be used as definition for a channel to be deterministic, see Exercise 2.5.12.

2.5.3 Joint states and channels

For an ordinary function $f \colon X \to Y$ one can form the *graph* of f as the relation $\mathit{gr}(f) \subseteq X \times Y$ given by $\mathit{gr}(f) = \{(x, y) \mid f(x) = y\}$. There is a similar thing that one can do for probabilistic channels, given a state. But please keep in mind that this kind of 'graph' has nothing to do with the graphical models that we will be using later, like Bayesian networks or string diagrams.

Definition 2.5.6. 1 For two channels $c\colon X \to\!\!\!\!\bullet\; Y$ and $d\colon X \to\!\!\!\!\bullet\; Z$ with the same domain we can define a tuple channel:

$$\langle c, d\rangle \coloneqq (c \otimes d) \mathbin{\circ} \Delta :\; X \to\!\!\!\!\bullet\; Y \times Z.$$

2 In particular, for a state σ on Y and a channel $d\colon Y \to Z$ we define the *graph* as the joint state on $Y \times Z$ defined by:

$$\langle id\,, d\rangle \gg\!= \sigma \qquad \text{so that} \qquad \Big(\langle id\,, d\rangle \gg\!= \sigma\Big)(y, z) \;=\; \sigma(y) \cdot d(y)(z).$$

We are overloading the tuple notation $\langle c, d\rangle$. Above we use it for the tuple of *channels*, so that $\langle c, d\rangle$ has type $X \to \mathcal{D}(Y \times Z)$. Interpreting $\langle c, d\rangle$ as a tuple of *functions*, like in Subsection 1.1.1, would give a type $X \to \mathcal{D}(Y) \times \mathcal{D}(Z)$. For channels we standardly use the channel-interpretation for the tuple.

In the literature a probabilistic channel is often called a *discriminative model*. Let's consider a channel whose codomain is a finite set, say (isomorphic to) the set $n = \{0, 1, \ldots, n-1\}$, whose elements $i \in n$ can be seen as labels or classes. Hence we can write the channel as $c\colon X \to\!\!\!\!\bullet\; n$. It can then be understood as a *classifier*: it produces for each element $x \in X$ a distribution $c(x)$ on n, giving the likelihood $c(x)(i)$ that x belongs to class $i \in n$. The class i with the highest likelihood, obtained via argmax, may simply be used as x's class.

The term *generative model* is used for a pair of a state $\sigma \in \mathcal{D}(Y)$ and a channel $g\colon Y \to\!\!\!\!\bullet\; Z$, giving rise to a joint state $\langle id\,, g\rangle \gg\!= \sigma \in \mathcal{D}(X \times Y)$. This shows how the joint state can be generated. Later on we shall see that in principle each joint state can be described in such generative form, via marginalisation and disintegration, see Section 7.6.

Exercises

2.5.1 1 Let $U \in \mathcal{P}(X)$ and $V \in \mathcal{P}(Y)$. Show that the equation

$$(U \otimes V)[1, 0] \;=\; U$$

does *not* hold in general, in particular not when $V = \emptyset$ (and $U \neq \emptyset$).

2 Similarly, check that for arbitrary multisets $\varphi \in \mathcal{M}(X)$ and $\psi \in \mathcal{M}(Y)$ one does *not* have:

$$(\varphi \otimes \psi)[0, 1] \;=\; \varphi.$$

It fails when ψ is but φ is not the empty multiset $\mathbf{0}$. But it also fails in non-zero cases, *e.g.* for the multisets $\varphi = 2|a\rangle + 4|b\rangle + 1|c\rangle$ and $\psi = 3|u\rangle + 2|v\rangle$. Check this by doing the required computations.

2.5.2 Check that the distribution $\omega \in \mathcal{D}(\{a,b\} \times \{a,b\} \times \{a,b\})$ given by:

$$\begin{aligned}\omega \;=\; \tfrac{1}{24}|aaa\rangle + \tfrac{1}{12}|aab\rangle + \tfrac{1}{12}|aba\rangle + \tfrac{1}{6}|abb\rangle \\ + \tfrac{1}{6}|baa\rangle + \tfrac{1}{3}|bab\rangle + \tfrac{1}{24}|bba\rangle + \tfrac{1}{12}|bbb\rangle\end{aligned}$$

satisfies:

$$\omega \;=\; \omega[1,1,0] \otimes \omega[0,0,1].$$

2.5.3 Check that for finite sets X, Y, the uniform distribution on $X \times Y$ is non-entwined — see also Exercise 2.1.1.

2.5.4 Find different joint states σ, τ with equal marginals:

$$\sigma[1,0] \;=\; \tau[1,0] \quad \text{and} \quad \sigma[0,1] \;=\; \tau[0,1] \quad \text{but} \quad \sigma \;\neq\; \tau.$$

Hint: Use Example 2.5.4.

2.5.5 Check that:

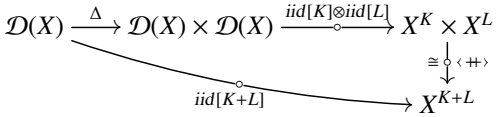
$$\omega[0,1,1,0,1,1][0,1,1,0] \;=\; \omega[0,0,1,0,1,0]$$

What is the general result behind this?

2.5.6 Show that a (binary) joint distribution is non-entwined when its first marginal is a unit (Dirac) distribution.

Formulated more explicitly, if $\omega \in \mathcal{D}(X \times Y)$ satisfies $\omega[1,0] = 1|x\rangle$, then $\omega \;=\; 1|x\rangle \otimes \omega[0,1]$. (This property shows that $\mathcal{D}$ is a 'strongly affine monad', see [69] for details.)

2.5.7 Prove that the following diagram about *iid* and concatenation ++ commutes, for $K, L \in \mathbb{N}$.



2.5.8 Prove that for a channel c with suitably typed codomain,

$$c[1,0,1,0,1] \;=\; \langle \pi_1, \pi_3, \pi_5 \rangle \mathbin{\circ} c.$$

2.5.9 Let $X = \{u, v\}$ and $A = \{a, b\}$ as in Example 2.5.4. Prove that a state $\omega = r_1|u,a\rangle + r_2|u,b\rangle + r_3|v,a\rangle + r_4|v,b\rangle \in \mathcal{D}(X \times A)$, where $r_1 + r_2 + r_3 + r_4 = 1$, is non-entwined if and only if $r_1 \cdot r_4 = r_2 \cdot r_3$.

2.5.10 Consider the probabilistic channel $f\colon X \nrightarrow Y$ from Example 2.2.1 and show that on the one hand $\Delta \mathbin{\circ} f\colon X \nrightarrow Y \times Y$ is given by:

$$\begin{aligned} a &\longmapsto \tfrac{1}{2}|u,u\rangle + \tfrac{1}{2}|v,v\rangle \\ b &\longmapsto 1|u,u\rangle \\ c &\longmapsto \tfrac{3}{4}|u,u\rangle + \tfrac{1}{4}|v,v\rangle. \end{aligned}$$

On the other hand, $(f \otimes f) \circ \Delta \colon X \nrightarrow Y \times Y$ is described by:

$$\begin{aligned} a &\longmapsto \tfrac{1}{4}|u,u\rangle + \tfrac{1}{4}|u,v\rangle + \tfrac{1}{4}|v,u\rangle + \tfrac{1}{4}|v,v\rangle \\ b &\longmapsto 1|u,u\rangle \\ c &\longmapsto \tfrac{9}{16}|u,u\rangle + \tfrac{3}{16}|u,v\rangle + \tfrac{3}{16}|v,u\rangle + \tfrac{1}{16}|v,v\rangle. \end{aligned}$$

2.5.11 Check that for ordinary functions $f \colon X \to Y$ and $g \colon X \to Z$ one can relate channel tupling and ordinary tupling via the operation $\langle - \rangle$ as:

$$\langle \langle f \rangle, \langle g \rangle \rangle = \langle \langle f, g \rangle \rangle \qquad \textit{i.e.} \qquad \langle \mathit{unit} \circ f, \mathit{unit} \circ g \rangle = \mathit{unit} \circ \langle f, g \rangle.$$

Conclude that the copy channel Δ is $\langle \mathit{unit}, \mathit{unit} \rangle$ and show that it does commute with *deterministic* channels:

$$\Delta \circledcirc \langle f \rangle = \langle \langle f \rangle, \langle f \rangle \rangle \circledcirc \Delta.$$

2.5.12 Deterministic channels are in fact the only ones that commute with copy (see for an abstract setting [18]). Let $c \colon X \nrightarrow Y$ commute with diagonals, in the sense that $\Delta \circledcirc c = (c \otimes c) \circledcirc \Delta$. Prove that c is deterministic, *i.e.* of the form $c = \langle f \rangle = \mathit{unit} \circ f$ for a function $f \colon X \nrightarrow Y$.

2.5.13 Check that the fact that in general $\Delta \gg\!= \omega \neq \omega \otimes \omega$, as described in Subsection 2.5.2, can also be expressed as: the following diagram does *not* commute:

$$\begin{array}{ccc} \mathcal{D}(X) & \xrightarrow{\quad \mathcal{D}(\Delta) \quad} & \mathcal{D}(X \times X) \\ & {\scriptstyle\Delta}\searrow \quad \nearrow{\scriptstyle\otimes} & \\ & \mathcal{D}(X) \times \mathcal{D}(X) & \end{array}$$

2.5.14 Show that the tuple operation $\langle -, - \rangle$ for channels, from Definition 2.5.6, satisfies both:

$$\pi_i \circledcirc \langle c_1, c_2 \rangle = c_i \qquad \text{and} \qquad \langle \pi_1, \pi_2 \rangle = \mathit{unit}$$

but *not*, essentially as in Exercise 2.5.10:

$$\langle c_1, c_2 \rangle \circledcirc d = \langle c_1 \circledcirc d, c_2 \circledcirc d \rangle.$$

2.5.15 Consider the joint state $\mathit{Flrn}(\tau)$ from Example 2.3.1 (2).

1 Take $\sigma = \frac{7}{10}|H\rangle + \frac{3}{10}|L\rangle \in \mathcal{D}(\{H, L\})$ and $c \colon \{H, L\} \nrightarrow \{0, 1, 2\}$ defined by:

$$c(H) = \tfrac{1}{7}|0\rangle + \tfrac{1}{2}|1\rangle + \tfrac{5}{14}|2\rangle \qquad c(T) = \tfrac{1}{6}|0\rangle + \tfrac{1}{3}|1\rangle + \tfrac{1}{2}|2\rangle.$$

Check that $\mathit{gr}(\sigma, c) = \mathit{Flrn}(\tau)$.

2.5.16 Consider a state $\sigma \in \mathcal{D}(X)$ and an 'endo' channel $c \colon X \nrightarrow X$.

1 Check that the following two statements are equivalent.

- $\langle id, c\rangle \gg\!= \sigma = \langle c, id\rangle \gg\!= \sigma$ in $\mathcal{D}(X \times X)$;
- $\sigma(x) \cdot c(x)(y) = \sigma(y) \cdot c(y)(x)$ for all $x, y \in X$.

2 Show that the conditions in the previous item imply that σ is a fixed point for state transformation, that is:

$$c \gg\!= \sigma = \sigma.$$

Such a σ is also called a stationary state.

2.5.17 Consider the bi-channel $bell\colon A\times 2 \nrightarrow B\times 2$ from Exercise 2.2.11 (2). Show that both $bell[1,0]$ and $bell^\dagger[1,0]$ are channels with uniform states only.

2.5.18 Prove that marginalisation $(-)[1,0]\colon \mathcal{M}(X \times Y) \to \mathcal{M}(X)$ is linear — and preserves convex sums when restricted to distributions.

2.6 A Bayesian network example

This section uses the previous first descriptions of probabilistic channels to give semantics for a Bayesian network example. At this stage we just describe an example, without giving the general approach, but we hope that it clarifies what is going on and illustrates to the reader the relevance of (probabilistic) channels and their operations. A general description of what a Bayesian network is appears much later, in Definition 7.1.2.

The example that we use is a standard one, copied from the literature, namely from [33]. Bayesian networks were introduced in [133], see also [109, 8, 13, 14, 91, 105]. They form a popular technique for displaying probabilistic connections and for efficiently presenting joint states, without state explosion.

Consider the diagram/network in Figure 2.3. It is meant to capture probabilistic dependencies between several wetness phenomena in the oval boxes. For instance, in winter it is more likely to rain (than when it's not winter), and also in winter it is less likely that a sprinkler is on. Still the grass may be wet by a combination of these occurrences. Whether a road is slippery depends on rain, not on sprinklers.

The letters A, B, C, D, E in this diagram are written exactly as in [33]. Here they are not used as sets of Booleans, with inhabitants *true* and *false*, but instead we use these sets with elements:

$$A = \{a, a^\perp\} \qquad B = \{b, b^\perp\} \qquad C = \{c, c^\perp\} \qquad D = \{d, d^\perp\} \qquad E = \{e, e^\perp\}.$$

The notation $a^\perp$ is read as 'not a'. In this way the name of an elements suggests to which set the element belongs.

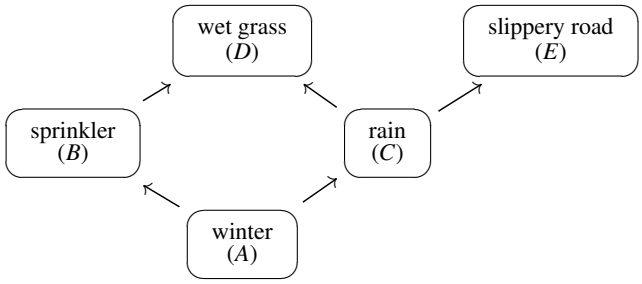


Figure 2.3 The wetness Bayesian network (from [33, Chap. 6]), with only the nodes and edges between them; the conditional probability tables associated with the nodes are given separately in the text.

The diagram in Figure 2.3 becomes a Bayesian network when we provide it with *conditional probability tables*. For the lower three nodes they look as follows.

winter	a	$a^\perp$
	$3/5$	$2/5$

sprinkler	A	b	$b^\perp$
	a	$1/5$	$4/5$
	$a^\perp$	$3/4$	$1/4$

rain	A	c	$c^\perp$
	a	$4/5$	$1/5$
	$a^\perp$	$1/10$	$9/10$

And for the upper two nodes we have:

wet grass	B	C	d	$d^\perp$
	b	c	$19/20$	$1/20$
	b	$c^\perp$	$9/10$	$1/10$
	$b^\perp$	c	$4/5$	$1/5$
	$b^\perp$	$c^\perp$	0	1

slippery road	C	e	$e^\perp$
	c	$7/10$	$3/10$
	$c^\perp$	0	1

This Bayesian network is thus given by nodes, each with a conditional probability table, describing likelihoods in terms of previous 'ancestor' nodes in the network (if any).

How to interpret all this data? How to make it mathematically precise? It is not hard to see that the first 'winter' table describes a probability distribution on the set A, which, in the notation of this book, is given by:

$$wi \;=\; \tfrac{3}{5}|a\rangle + \tfrac{2}{5}|a^\perp\rangle.$$

Thus we are assuming with probability of 60% that we are in a winter situation. This is often called the *prior* distribution, or also the *initial* state.

Notice that the 'sprinkler' table contains two distributions on B, one for the

element $a \in A$ and one for $a^\perp \in A$. Here we recognise a channel, namely a channel $A \multimap B$. This is a crucial insight! We abbreviate this channel as sp, and define it explicitly as:

$$A \xrightarrow{sp} B \qquad \text{with} \qquad \begin{cases} sp(a) = \frac{1}{5}|b\rangle + \frac{4}{5}|b^\perp\rangle \\ sp(a^\perp) = \frac{3}{4}|b\rangle + \frac{1}{4}|b^\perp\rangle. \end{cases}$$

We read them as: if it's winter, there is a 20% chance that the sprinkler is on, but if it's not winter, there is 75% chance that the sprinkler is on.

Similarly, the 'rain' table corresponds to a channel:

$$A \xrightarrow{ra} C \qquad \text{with} \qquad \begin{cases} ra(a) = \frac{4}{5}|c\rangle + \frac{1}{5}|c^\perp\rangle \\ ra(a^\perp) = \frac{1}{10}|c\rangle + \frac{9}{10}|c^\perp\rangle. \end{cases}$$

Before continuing we can see that the formalisation (partial, so far) of the wetness Bayesian network in Figure 2.3 in terms of states and channels, already allows us to do something meaningful, namely state transformation $\gg=$. Indeed, we can form distributions:

$$sp \gg= wi \ \text{ on } B \qquad \text{and} \qquad ra \gg= wi \ \text{ on } C.$$

These (transformed) distributions capture the derived, predicted probabilities that the sprinkler is on, and that it rains. Using the definition of state transformation, see (2.9), we get:

$$\begin{aligned} (sp \gg= wi)(b) &= \textstyle\sum_x sp(x)(b) \cdot wi(x) \\ &= sp(a)(b) \cdot wi(a) + sp(a^\perp)(b) \cdot wi(a^\perp) \\ &= \tfrac{1}{5} \cdot \tfrac{3}{5} + \tfrac{3}{4} \cdot \tfrac{2}{5} = \tfrac{21}{50} \\ (sp \gg= wi)(b^\perp) &= \textstyle\sum_x sp(x)(b^\perp) \cdot wi(x) \\ &= sp(a)(b^\perp) \cdot wi(a) + sp(a^\perp)(b^\perp) \cdot wi(a^\perp) \\ &= \tfrac{4}{5} \cdot \tfrac{3}{5} + \tfrac{1}{4} \cdot \tfrac{2}{5} = \tfrac{29}{50}. \end{aligned}$$

Thus the overall distribution for the sprinkler (being on or not) is:

$$sp \gg= wi = \tfrac{21}{50}|b\rangle + \tfrac{29}{50}|b^\perp\rangle.$$

In a similar way one can compute the probability distribution for rain as:

$$ra \gg= wi = \tfrac{13}{25}|c\rangle + \tfrac{12}{25}|c^\perp\rangle.$$

Such distributions for non-initial nodes of a Bayesian network are called *predictions*. As will be shown here, they can be obtained via forward state transformation, following the structure of the network.

But first we still have to translate the upper two nodes of the network from Figure 2.3 into channels. In the conditional probability table for the 'wet grass'

node we see 4 distributions on the set D, one for each combination of elements from the sets B and C. The table thus corresponds to a channel:

$$B \times C \xrightarrow{wg} D \qquad \text{with} \qquad \begin{cases} wg(b, c) = \frac{19}{20}|d\rangle + \frac{1}{20}|d^\perp\rangle \\ wg(b, c^\perp) = \frac{9}{10}|d\rangle + \frac{1}{10}|d^\perp\rangle \\ wg(b^\perp, c) = \frac{4}{5}|d\rangle + \frac{1}{5}|d^\perp\rangle \\ wg(b^\perp, c^\perp) = 1|d^\perp\rangle. \end{cases}$$

Finally, the table for the 'slippery road' table gives:

$$C \xrightarrow{sr} E \qquad \text{with} \qquad \begin{cases} sr(c) = \frac{7}{10}|e\rangle + \frac{3}{10}|e^\perp\rangle \\ sr(c^\perp) = 1|e^\perp\rangle. \end{cases}$$

We illustrate how to obtain predictions for 'rain' and for 'slippery road'. We start from the latter. Looking at the network in Figure 2.3 we see that there are two arrows between the initial node 'winter' and our node of interest 'slippery road'. This means that we have to do two state successive transformations, giving:

$$\begin{aligned} (sr \circ ra) \gg\!= wi &= sr \gg\!= (ra \gg\!= wi) \\ &= sr \gg\!= (\tfrac{13}{25}|c\rangle + \tfrac{12}{25}|c^\perp\rangle) = \tfrac{91}{250}|e\rangle + \tfrac{159}{250}|e^\perp\rangle. \end{aligned}$$

The first equation follows from Lemma 1.8.3 (3). The second one involves elementary calculations, where we can use the distribution $ra \gg\!= wi$ that we calculated earlier.

Getting the predicted wet grass probability requires some care. Inspection of the network in Figure 2.3 is of some help, but leads to some ambiguity — see below. One might be tempted to form the parallel product $\otimes$ of the predicted distributions for sprinkler and rain, and do state transformation on this product along the wet grass channel wg, as in:

$$wg \gg\!= ((sp \gg\!= wi) \otimes (ra \gg\!= wi)).$$

But this is wrong, since the winter probabilities are now not used consistently, see the different outcomes in the calculations (2.24) and (2.25). The correct way to obtain the wet grass prediction involves copying the winter state, via the copy channel Δ, see:

$$\begin{aligned} (wg \circ (sp \otimes ra) \circ \Delta) \gg\!= wi &= wg \gg\!= \Big((sp \otimes ra) \gg\!= (\Delta \gg\!= wi)\Big) \\ &= wg \gg\!= (\langle sp, ra\rangle \gg\!= wi) \\ &= \tfrac{1399}{2000}|d\rangle + \tfrac{601}{2000}|d^\perp\rangle. \end{aligned}$$

Such calcutions are laborious, but essentially straightforward. We shall do this in one in detail, just to see how it works. Especially, it becomes clear that all

summations are automatically done at the right place. We proceed in two steps, where for each step we only elaborate the first case.

$$\begin{array}{rcl}
(\langle sp, ra\rangle \gg\!= wi)(b, c) &=& \sum_x sp(x)(b) \cdot ra(x)(c) \cdot wi(x) \\
&=& sp(a)(b) \cdot ra(a)(c) \cdot \frac{3}{5} + sp(a^\perp)(b) \cdot ra(a^\perp)(c) \cdot \frac{2}{5} \\
&=& \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} + \frac{3}{4} \cdot \frac{1}{10} \cdot \frac{2}{5} \;=\; \frac{63}{500} \\
(\langle sp, ra\rangle \gg\!= wi)(b, c^\perp) &=& \cdots \;=\; \frac{147}{500} \\
(\langle sp, ra\rangle \gg\!= wi)(b^\perp, c) &=& \frac{197}{500} \\
(\langle sp, ra\rangle \gg\!= wi)(b^\perp, c^\perp) &=& \frac{93}{500}.
\end{array}$$

We conclude that:

$$\begin{array}{rcl}
(sp \otimes ra) \gg\!= (\Delta \gg\!= wi) &=& \langle sp, ra\rangle \gg\!= wi \\
&=& \frac{63}{500}|b, c\rangle + \frac{147}{500}|b, c^\perp\rangle + \frac{197}{500}|b^\perp, c\rangle + \frac{93}{500}|b^\perp, c^\perp\rangle.
\end{array}$$

This distribution is used in the next step:

$$\begin{array}{l}
\Big(wg \gg\!= \big((sp \otimes ra) \gg\!= (\Delta \gg\!= wi)\big)\Big)(d) \\
\quad= \sum_{x,y} wg(x, y)(d) \cdot \big((sp \otimes ra) \gg\!= (\Delta \gg\!= wi)\big)(x, y) \\
\quad= wg(b, c)(d) \cdot \frac{63}{500} + wg(b, c^\perp)(d) \cdot \frac{147}{500} \\
\qquad\quad + wg(b^\perp, c)(d) \cdot \frac{197}{500} + wg(b^\perp, c^\perp)(d) \cdot \frac{93}{500} \\
\quad= \frac{19}{20} \cdot \frac{63}{500} + \frac{9}{10} \cdot \frac{147}{500} + \frac{4}{5} \cdot \frac{197}{500} + 0 \cdot \frac{93}{500} \\
\quad= \frac{1399}{2000} \\
\Big(wg \gg\!= \big((sp \otimes ra) \gg\!= (\Delta \gg\!= wi)\big)\Big)(d^\perp) \\
\quad= \frac{601}{2000}.
\end{array}$$

We have thus shown that:

$$wg \gg\!= \big((sp \otimes ra) \gg\!= (\Delta \gg\!= wi)\big) \;=\; \frac{1399}{2000}|d\rangle + \frac{601}{2000}|d^\perp\rangle.$$

2.6.1 Redrawing Bayesian networks

We have illustrated how prediction computations for Bayesian networks can be done, basically by following the graph structure and translating it into suitable sequential and parallel compositions ($\circ$ and $\otimes$) of channels. The match between the graph and the computation is not perfect, and requires some care, especially wrt. copying. Since we have a solid semantics, we like to use it in order to improve the network drawing and achieve a better match between the underlying mathematical operations and the graphical representation. Therefore we prefer to draw Bayesian networks slightly differently, making some minor changes:

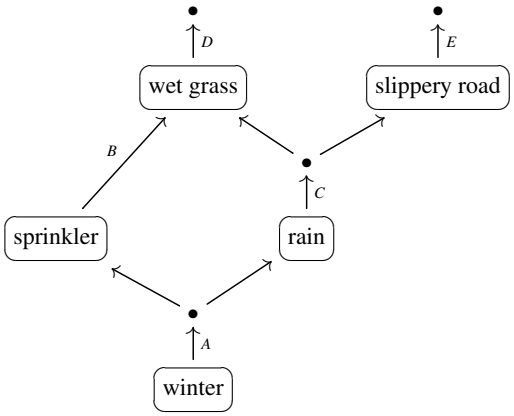


Figure 2.4 The wetness Bayesian network from Figure 2.3 redrawn in a manner that better reflects the underlying channel-based semantics, via explicit copiers and typed wires. This anticipates the string-diagrammatic approach of Chapter 7.

- copying is written explicity, for instance as $\curlyvee$ in the binary case; in general one can have n-ary copying, for $n \geq 2$;
- the relevant sets/types — like A, B, C, D, E — are not included in the nodes, but are associated with the arrows (wires) between the nodes;
- final nodes have outgoing arrows, labeled with their type.

The original Bayesian network in Figure 2.3 is then changed according to these points in Figure 2.4. In this way the nodes (with their conditional probability tables) are clearly recognisable as channels, of type $A_1 \times \cdots \times A_n \mathrel{-\!\!\circ\!\!\to} B$, where $A_1, \ldots, A_n$ are the types on the incoming wires, and B is the type of the outgoing wire. Initial nodes have no incoming wires, which formally leads to a channel $1 \mathrel{-\!\!\circ\!\!\to} B$, where 1 is the empty product. As we have seen, such channels $1 \mathrel{-\!\!\circ\!\!\to} B$ correspond to distributions/states on B. In the adapted diagram one easily forms sequential and parallel compositions of channels, see the exercise below.

Exercises

2.6.1 In [33, §6.2] the (predicted) joint distribution on $D \times E$ that arises from the Bayesian network example in this section is reprented as a table. It translates into:

$$\tfrac{30{,}443}{100{,}000}|d, e\rangle + \tfrac{39{,}507}{100{,}000}|d, e^\perp\rangle + \tfrac{5{,}957}{100{,}000}|d^\perp, e\rangle + \tfrac{24{,}093}{100{,}000}|d^\perp, e^\perp\rangle.$$

Following the structure of the diagram in Figure 2.4, it is obtained in the present setting as:

$$\begin{aligned}&\Big((wg \otimes sr) \circ (id \otimes \Delta) \circ (sp \otimes ra) \circ \Delta\Big) \gg\!= wi \\ &= (wg \otimes sr) \gg\!= ((id \otimes \Delta) \gg\!= (\langle sp, ra\rangle \gg\!= wi)).\end{aligned}$$

Perform the calculations and check that this expression equals the above distribution.

(Readers may wish to compare the different calculation methods, using sequential and parallel composition of channels — as here — or using multiplications of tables — as in [33].)

2.7 Divergence between distributions

In many situations it is useful to know how unequal / different / apart probability distributions are. This can be used for instance in learning, where one can try to bring a distribution closer to target via iterative adaptations. Such comparison of distributions can be defined via a metric / distance function. Here we shall describe a different comparison, called *divergence*, or more fully *Kullback-Leibler divergence*, written as D_{KL}. It is a not a distance function since it is not symmetric: $D_{KL}(\omega, \rho) \neq D_{KL}(\rho, \omega)$, in general. But it does satisfy $D_{KL}(\omega, \rho) = 0$ iff $\omega = \rho$. In Section 4.5 we describe 'total variation' as a proper distance function between states.

In the sequel we shall make frequent use of this divergence. This section collects the definition and some basic facts. It assumes rudimentary familiarity with logarithms.

Definition 2.7.1. Let ω, ρ be two distributions/states on the same set X with $supp(\omega) \subseteq supp(\rho)$. The *Kullback-Leibler divergence*, or *KL-divergence*, or simply *divergence*, of ω from ρ is:

$$D_{KL}(\omega, \rho) \coloneqq \sum_{x \in X} \omega(x) \cdot \log\left(\frac{\omega(x)}{\rho(x)}\right).$$

The convention is that $r \cdot \log(r) = 0$ when $r = 0$.

Sometimes the natural logarithm ln is used instead of $\log = \log_2$.

The inclusion $supp(\omega) \subseteq supp(\rho)$ is equivalent to: $\rho(x) = 0$ implies $\omega(x) = 0$. This requirement immediately implies that divergence is not symmetric. But even when ω and ρ do have the same support, the divergences $D_{KL}(\omega, \rho)$ and $D_{KL}(\rho, \omega)$ are different, in general, see Exercise 2.7.1 below for an easy illustration.

Whenever we write an expression $D_{KL}(\omega, \rho)$ we will implicitly assume an inclusion $supp(\omega) \subseteq supp(\rho)$.

We start with some easy properties of divergence.

Lemma 2.7.2. *Let $\omega, \rho \in \mathcal{D}(X)$ and $\omega', \rho' \in \mathcal{D}(Y)$ be distributions.*

1 Zero-divergence is the same as equality:

$$D_{KL}(\omega, \rho) = 0 \iff \omega = \rho.$$

2 Divergence of products is a sum of divergences:

$$D_{KL}(\omega \otimes \omega',\, \rho \otimes \rho') \;=\; D_{KL}(\omega, \rho) + D_{KL}(\omega', \rho').$$

Proof. 1 The direction ($\Leftarrow$) is easy. For ($\Rightarrow$), let $0 = D_{KL}(\omega, \rho) = \sum_x \omega(x) \cdot \log\left({}^{\omega(x)}/_{\rho(x)}\right)$. This means that if $\omega(x) \neq 0$, one has $\log\left({}^{\omega(x)}/_{\rho(x)}\right) = 0$, and thus ${}^{\omega(x)}/_{\rho(x)} = 1$ and $\omega(x) = \rho(x)$. In particular:

$$1 \;=\; \sum_{x \in supp(\omega)} \omega(x) \;=\; \sum_{x \in supp(\omega)} \rho(x). \tag{$*$}$$

By assumption we have $supp(\omega) \subseteq supp(\rho)$. Write $supp(\rho)$ as disjoint union $supp(\omega) \cup U$ for some $U \subseteq supp(\rho)$. It suffices to show $U = \emptyset$. We have:

$$1 \;=\; \sum_{x \in supp(\rho)} \rho(x) \;=\; \sum_{x \in supp(\omega)} \rho(x) + \sum_{x \in U} \rho(x) \overset{(*)}{=} 1 + \sum_{x \in U} \rho(x).$$

Hence $U = \emptyset$.

2 By unwrapping the relevant definitions:

$$\begin{aligned}
&D_{KL}(\omega \otimes \omega',\, \rho \otimes \rho') \\
&= \sum_{x,y} (\omega \otimes \omega')(x, y) \cdot \log\left(\frac{(\omega \otimes \omega')(x, y)}{(\rho \otimes \rho')(x, y)}\right) \\
&= \sum_{x,y} \omega(x) \cdot \omega'(y) \cdot \log\left(\frac{\omega(x)}{\rho(x)} \cdot \frac{\omega'(y)}{\rho'(y)}\right) \\
&= \sum_{x,y} \omega(x) \cdot \omega'(y) \cdot \left(\log\left(\frac{\omega(x)}{\rho(x)}\right) + \log\left(\frac{\omega'(y)}{\rho'(y)}\right)\right) \\
&= \sum_{x,y} \omega(x) \cdot \omega'(y) \cdot \log\left(\frac{\omega(x)}{\rho(x)}\right) \;+\; \sum_{x,y} \omega(x) \cdot \omega'(y) \cdot \log\left(\frac{\omega'(y)}{\rho'(y)}\right) \\
&= \sum_{x} \omega(x) \cdot \log\left(\frac{\omega(x)}{\rho(x)}\right) \;+\; \sum_{y} \omega'(y) \cdot \log\left(\frac{\omega'(y)}{\rho'(y)}\right) \\
&= D_{KL}(\omega, \rho) + D_{KL}(\omega', \rho').
\end{aligned}$$

❑

In order to prove further properties about divergence we need a powerful classical result called *Jensen's inequality* about functions acting on convex combinations of non-negative reals. We shall use Jensen's inequality below, but also later on in learning.

Lemma 2.7.3 (Jensen's inequality)**.** *Let $f\colon \mathbb{R}_{>0} \to \mathbb{R}$ be a function that satisfies $f'' < 0$. Then for all $a_1, \ldots, a_n \in \mathbb{R}_{>0}$ and $r_1, \ldots, r_n \in [0,1]$ with $\sum_i r_i = 1$ there is an inequality:*

$$f\left(\sum\nolimits_i r_i \cdot a_i\right) \geq \sum\nolimits_i r_i \cdot f(a_i). \tag{2.26}$$

The inequality is strict, except in trivial cases.

The inequality holds in particular for logarithms, when $f = \log$, or $f = \ln$.

The proof is standard but is included, for convenience.

Proof. We shall provide a proof for $n = 2$. The inequality is easily extended to $n > 2$, by induction. So let $a, b \in \mathbb{R}_{>0}$ be given, with $r \in [0,1]$. We need to prove $f(ra + (1-r)b) \geq rf(a) + (1-r)f(b)$. The result is trivial if $a = b$ or $r = 0$ or $r = 1$. So let, without loss of generality, $a < b$ and $r \in (0,1)$. Write $c := ra + (1-r)b = b - r(b-a)$, so that $a < c < b$. By the mean value theorem we can find $a < u < c$ and $c < v < b$ with:

$$\frac{f(c) - f(a)}{c - a} = f'(u) \qquad \text{and} \qquad \frac{f(b) - f(c)}{b - c} = f'(v)$$

Since $f'' < 0$ we have that f' is strictly decreasing, so $f'(u) > f'(v)$ because $u > v$. We can write:

$$c - a = (r-1)a + (1-r)b = (1-r)(b-a) \quad \text{and} \quad b - c = r(b-a).$$

From $f'(u) > f'(v)$ we deduce inequalities:

$$\frac{f(c) - f(a)}{(1-r)(b-a)} > \frac{f(b) - f(c)}{r(b-a)} \quad \textit{i.e.} \quad r(f(c) - f(a)) > (1-r)(f(b) - f(c)).$$

By reorganising the latter inequality we get $f(c) > rf(a) + (1-r)f(b)$, as required. ❑

We can now say a bit more about divergence. For instance, that it is non-negative, as one expects.

Proposition 2.7.4. *Let $\omega, \rho \in \mathcal{D}(X)$ be states on the same space X.*

1. *$D_{KL}(\omega, \rho) \geq 0$.*
2. *State transformation is D_{KL}-non-expansive: for a channel $c\colon X \to\!\!\!\!\bullet\; Y$ and states $\omega, \rho \in \mathcal{D}(X)$ one has:*

$$D_{KL}(c \gg= \omega,\ c \gg= \rho) \leq D_{KL}(\omega, \rho).$$

Proof. 1 Via Jensen's inequality, see we get:

$$\begin{aligned} -D_{KL}(\omega,\rho) &= \sum\nolimits_x \omega(x)\cdot\log\left(\frac{\rho(x)}{\omega(x)}\right) \\ &\leq \log\left(\sum\nolimits_x \omega(x)\cdot\frac{\rho(x)}{\omega(x)}\right) = \log\left(\sum\nolimits_x \rho(x)\right) = \log(1) = 0. \end{aligned}$$

2 Again Via Jensen's inequality:

$$\begin{aligned} D_{KL}(c \gg\!= \omega, c \gg\!= \rho) &= \sum\nolimits_y (c \gg\!= \omega)(y)\cdot\log\left(\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\right) \\ &= \sum\nolimits_{x,y} \omega(x)\cdot c(x)(y)\cdot\log\left(\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\right) \\ &\leq \sum\nolimits_x \omega(x)\cdot\log\left(\sum\nolimits_y c(x)(y)\cdot\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\right). \end{aligned}$$

Hence it suffices to prove:

$$\begin{aligned} \sum\nolimits_x \omega(x)\cdot\log\left(\sum\nolimits_y c(x)(y)\cdot\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\right) &\leq D_{KL}(\omega,\rho) \\ &= \sum\nolimits_x \omega(x)\cdot\log\left(\frac{\omega(x)}{\rho(x)}\right). \end{aligned}$$

This inequality $\leq$ follows from another application of Jensen's inequality:

$$\begin{aligned} &\sum\nolimits_x \omega(x)\cdot\log\left(\sum\nolimits_y c(x)(y)\cdot\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\right) - \sum\nolimits_x \omega(x)\cdot\log\left(\frac{\omega(x)}{\rho(x)}\right) \\ &= \sum\nolimits_x \omega(x)\cdot\left[\log\left(\sum\nolimits_y c(x)(y)\cdot\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\right) - \log\left(\frac{\omega(x)}{\rho(x)}\right)\right] \\ &= \sum\nolimits_x \omega(x)\cdot\log\left(\sum\nolimits_y c(x)(y)\cdot\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\cdot\frac{\rho(x)}{\omega(x)}\right) \\ &\leq \log\left(\sum\nolimits_{x,y} \omega(x)\cdot c(x)(y)\cdot\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\cdot\frac{\rho(x)}{\omega(x)}\right) \\ &= \log\left(\sum\nolimits_y \left(\sum\nolimits_x c(x)(y)\cdot\rho(x)\right)\cdot\frac{(c \gg\!= \omega)(y)}{(c \gg\!= \rho)(y)}\right) \\ &= \log\left(\sum\nolimits_y (c \gg\!= \omega)(y)\right) \\ &= \log(1) = 0. \end{aligned}$$

❑

Exercises

2.7.1 Take $\omega = \frac{1}{4}|a\rangle + \frac{3}{4}|b\rangle$ and $\rho = \frac{1}{2}|a\rangle + \frac{1}{2}|b\rangle$. Check that:

1 $D_{KL}(\omega, \rho) = \frac{3}{4} \cdot \log(3) - 1 \approx 0.19$.
2 $D_{KL}(\rho, \omega) = 1 - \frac{1}{2} \cdot \log(3) \approx 0.21$.

2.7.2 Check that for $y \in \mathit{supp}(\rho)$ one has:

$$D_{KL}(1|y\rangle, \rho) \;=\; -\log\left(\rho(y)\right).$$

2.7.3 Show that the function $D_{KL}(\omega, -)$ behaves as follows on convex combinations of states:

$$\begin{aligned} &D_{KL}(\omega, r_1 \cdot \rho_1 + \cdots + r_n \cdot \rho_n) \\ &\leq\; r_1 \cdot D_{KL}(\omega, \rho_1) + \cdots + r_n \cdot D_{KL}(\omega, \rho_n). \end{aligned}$$

2.7.4 Use Jensen's inequality to prove what is known as the inequality of arithmetic and geometric means: for $r_i, a_i \in \mathbb{R}_{\geq 0}$ with $\sum_i r_i = 1$,

$$\sum\nolimits_i r_i \cdot a_i \;\geq\; \prod\nolimits_i a_i^{r_i}.$$

3

Drawing from an urn

One of the basic topics covered in textbooks on probability is drawing from an urn, see *e.g.* [139, 145, 115] and many other references. An urn is container which is typically filled with balls of different colours. The idea is that the balls in the urn are thoroughly shuffled so that the urn forms a physical representation of a probability distribution. The challenge is to describe the probabilities associated with the experiment of (blindly) drawing one or more balls from the urn and registering the colour(s). Commonly two questions can be asked in this setting.

- Do we consider the drawn balls in order, or not? More specifically, do we see a draw of multiple balls as a list, or as a multiset?
- What happens to the urn after a ball is drawn? We consider three scenarios, with short names "-1", "0", "+1", see Examples 2.1.1.
  - In the "-1" scenario the drawn ball is removed from the urn. This is covered by the hypergeometric distribution.
  - In the "0" scenario the drawn ball is returned to the urn, so that the urn remains unchanged. In that case the multinomial distribution applies.
  - In the "+1" scenario not only the drawn ball is returned to the urn, but one additional ball of the same colour is added to the urn. This is covered by the Pólya distribution.

  In the "-1" and "+1" scenarios the probability of drawing a ball of a certain colour changes after each draw. In the "-1" case only finitely many draws are possible, until the urn is empty.

This yields six variations in total, namely with ordered / unordered draws and with -1, 0, 1 as replacement scenario. These six options are represented in a 3×2 table, see (3.3) below.

An important goal of this chapter is to describe these six cases in a principled manner via probabilistic channels. When drawing from an urn one can concentrate on the draw — on what is in your hand after the draw — or on the urn — on what is left in the urn after the draw. It turns out to be most fruitful to combine these two perspectives in a single channel, acting as a transition map. This transition aspect will be studied more systematically in terms of probabilistic automata in Chapter **??**. Here we look at single (ball) draws, which may be described informally as:

$$\mathit{Urn} \longmapsto \big(\mathit{single\text{-}colour},\ \mathit{Urn}'\big) \tag{3.1}$$

We use the *ad hoc* notation Urn' to describe the urn after the draw. It may be the same urn as before, in case of a draw with replacement, or it may be a different urn, with one ball less/more, namely the original urn minus/plus a ball as drawn.

The above transition (3.1) will be described as a probabilistic channel. It gives for each single draw the associated probability. In this description we combine multisets and distributions. For instance, an urn with three red balls and two blue ones will be described as a (natural) multiset $3|R\rangle + 2|B\rangle$. The transition associated with drawing a single ball *without* replacement (scenario "-1") gives a mapping:

$$3|R\rangle + 2|B\rangle \longmapsto \tfrac{3}{5}\big|R,\ 2|R\rangle + 2|B\rangle\big\rangle + \tfrac{2}{5}\big|B,\ 3|R\rangle + 1|B\rangle\big\rangle.$$

It gives the $\frac{3}{5}$ probability of drawing a red ball, together with the remaining urn, and a $\frac{2}{5}$ probability of drawing a blue one, with a different new urn.

The "+1" scenario with *double* replacement gives a mapping:

$$3|R\rangle + 2|B\rangle \longmapsto \tfrac{3}{5}\big|R,\ 4|R\rangle + 2|B\rangle\big\rangle + \tfrac{2}{5}\big|B,\ 3|R\rangle + 3|B\rangle\big\rangle.$$

Finally, the situation with *single* replacement ("0") is given by:

$$3|R\rangle + 2|B\rangle \longmapsto \tfrac{3}{5}\big|R,\ 3|R\rangle + 2|B\rangle\big\rangle + \tfrac{2}{5}\big|B,\ 3|R\rangle + 2|B\rangle\big\rangle.$$

In this last, third case we see that the urn/multiset does not change. An important first observation is that in that case we may as well use a distribution as urn, instead of a multiset. The distribution represents an abstract urn. In the above example we would use the distribution $\frac{3}{5}|R\rangle + \frac{2}{5}|B\rangle$ as abstract urn, when we draw with single replacement (case "0"). The distribution contains all the relevant information. Clearly, it is obtained via frequentist learning from the original multiset. Using distributions instead of multisets gives more flexibility,

since not all distributions are obtained via frequentist learing — in particular when the probabilities are proper real numbers and not fractions.

We formulate this approach explicitly.

- In the "-1" and "+1" situations *without* or *with double* replacement, an urn is a (non-empty, natural) multiset, which changes with every draw, via removal or double replacement of the drawn ball. These scenarios will also be described in terms of deletion or addition, using the draw-delete and draw-add channels DD and DA that we already saw in Definition 2.2.4.
- In a "0" situation *with* replacement, an urn is a probability distribution; it does not change when balls are drawn.

This covers the above second question, about what happens to the urn. For the first question concerning ordered and unordered draws one has to go beyond single draw transitions. Hence we need to suitably iterate the single-draw transition (3.1) to:

$$\mathit{Urn} \longmapsto \Big(\mathit{multiple\text{-}colours},\ \mathit{Urn}'\Big) \tag{3.2}$$

Now we can make the distinction between ordered and unordered draws explicit. Let X be the set of colours, for the balls in the urn — so $X = \{R, B\}$ in the above illustration.

- An *ordered* draw of multiple balls, say K many, is represented via a list $X^K = X \times \cdots \times X$ of length K.
- An *unordered* draw of K-many balls is represented as a K-sized multiset, in $\mathcal{N}[K](X)$.

Thus, in the latter case, both the urn and the handful of balls drawn from it, are represented as a multiset.

In the end we are interested in assigning probabilities to draws, ordered or not, in "-1", "0" or "1" mode. These probabilities on draws are obtained by taking the first marginal/projection of the iterated transition map (3.2). It yields a mapping from an urn to multiple draws. The following table gives an overview of the types of these operations, where X is the set of colours of the balls.

mode	**ordered**	**unordered**
0	$\mathcal{D}(X) \xrightarrow{O_0} X^K$	$\mathcal{D}(X) \xrightarrow{U_0} \mathcal{N}[K](X)$
-1	$\mathcal{N}[L](X) \xrightarrow{O_-} X^K$	$\mathcal{N}[L](X) \xrightarrow{U_-} \mathcal{N}[K](X)$
+1	$\mathcal{N}_*(X) \xrightarrow{O_+} X^K$	$\mathcal{N}_*(X) \xrightarrow{U_+} \mathcal{N}[K](X)$

(3.3)

We see that in the replacement scenario "0" the inputs of these channels are distributions in $\mathcal{D}(X)$, as abstract urns. In the deletion scenario "-1" the input (urns) are multisets in $\mathcal{N}[L](X)$, of size L. In the ordered case the outputs are tuples in X^K of length K and in the unordered case they are multisets in $\mathcal{N}[K](X)$ of size K. Implicitly in this table we assume that $L \geq K$, so that the urn is full enough for K single draws. In the addition scenario "+1" we only require that the urn is a non-empty multiset, so that at least one ball can be drawn. Sometimes it is required that there is at least one ball of each colour in the urn, so that all colours can occur in draws.

The above table uses systematic names for the six different draw maps. In the unordered case the following (historical) names are common:

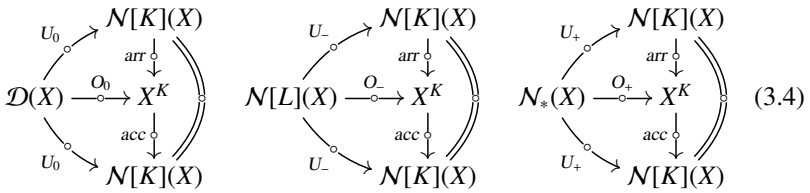
$$U_0 = \text{multinomial} \qquad U_- = \text{hypergeometric} \qquad U_+ = \text{Pólya}.$$

In this chapter we shall see that these three draw maps have certain properties in common, such as:

- Frequentist learning applied to the draws yields the same outcome as frequentist learning from the urn, see Theorem 3.3.5, Corollary 3.4.2 (2) and Proposition 3.4.5 (1).
- Doing a draw-delete DD after a draw of size $K+1$ is the same as doing a K-sized draw, see Proposition 3.3.8, Corollary 3.4.2 (4) and Theorem 3.4.6.

But we shall also see many differences between the three forms of drawing.

This chapter describes and analyses the six probabilistic channels in Table 3.3 for drawing from an urn. It turns out that many of the relevant properties can be expressed via composition of such channels, either sequentially (via $\circ$) or in parallel (via $\otimes$). In this analysis the operations of accumulation *acc* and arrangement *arr*, for going back-and-forth between products and multisets, play an important role. For instance, in each case one has commuting diagrams of the form.



(3.4)

Such commuting diagrams are unusual in the area of probability theory but we like to use them because they are very expressive. They involve multiple equations and clearly capture the types and order of the various operations involved.

Moreover, to emphasise once more, these are diagrams of *channels* with *channel composition* $\circledcirc$. As ordinary functions, with ordinary function composition $\circ$, there are no such diagrams / equations.

The chapter first takes a fresh look at accumulation and arrangement, introduced earlier in Sections 1.6 and 2.2. These are the operations for turning a list into a multiset, and a multiset into a uniform distribution of lists (that accumulate to the multiset). In Section 3.2 we will use accumulation and arrangement for the powerful operation for "zipping" two multisets of the same size together. It works analogously to zipping of two lists of the same length, into a single list of pairs, but this 'multizip' produces a distribution over (combined) multisets. The multizip operation turns out to interact smoothly with multinomial and hypergeometric distributions, as we shall see later on in this chapter.

Section 3.3 investigates multinomial channels and Section 3.4 covers both hypergeometric and Pólya channels, all in full, multivariate generality. We have briefly seen the multinomial and hypergeometric channels in Example 2.1.1. Now we take a much closer look, based on [78], and describe how these channels commute with basic operations such as accumulation and arrangement, with frequentist learning, with multizip, and with draw-and-delete. All these commutation properties involve channels and channel composition $\circledcirc$.

Subsequently, Section 3.5 elaborates how the channels in Table 3.3 actually arise. It makes the earlier informal descriptions in (3.1) and (3.2) mathematically precise. What happens is a bit sophisticated. We recall that for any monoid M, the mapping $X \mapsto M \times X$ is a monad, called the writer monad, see Lemma 1.9.1. This can be combined with the distribution monad $\mathcal{D}$, giving a combined monad $X \mapsto \mathcal{D}(M \times X)$. It comes with an associated 'Kleisli' composition. It is precisely this composition that we use for iterating a single draw, that is, for going from (3.1) to (3.2). Moreover, for ordered draws we use the monoid $M = \mathcal{L}(X)$ of lists, and for unordered draws we use the monoid $M = \mathcal{N}(X)$ of multisets. It is rewarding, from a formal perspective, to see that from this abstract principled approach, common distributions for different sorts of drawing arise, including the well known multinomial, hypergeometric and Pólya distributions. This is based on [81].

The subsequent two sections 3.6 and 3.7 of this chapter focus on a non-trivial operation from [78], namely turning a multiset of distributions into a distribution of multisets. Technically, this operation is a *distributive law*, called the parallel multinomial law. We spend ample time introducing it: Section 3.6 contains no less than four different definitions — all equivalent. Subsequently, various properties are demonstrated of this parallel multinomial law, including

commutation with hypergeometric channels, with frequentist learning and with muliset-zip.

The final section 3.9 briefly looks at yet another form of drawing coloured balls from an urn. The draws considered there are like Pólya urns — where an additional ball of the same colour as the drawn ball is returned to the urn — but with a new twist: after each draw another ball is added to the urn with a fresh colour, not already occurring in the urn. This gives an entirely new dynamics. The new colour works like a new mutation in genetics, and indeed, these kind of "Hoppe" urns have arisen in mathematical biology, first in the work of Ewens [46]. We describe the resulting distributions via probabilistic channels and show that there are some similarities and differences with the classical draws — multinomial, hypergeometric, and Pólya — where the number of colours of balls in the urn does not increase.

This chapter make more use of the language of category theory than other chapters in this book. It demonstrates that that there is ample (categorical) structure in basic probabilistic operations. The various properties of multinomial, hypergeometric and Pólya channels that occur in this chapter wil be used frequently in the sequel. Other, more technical topics could be skipped at first reading. An even more categorical, axiomatic approach can be found in [80].

3.1 Accumlation and arrangement, revisited

Accumulation and arrangement are the operations that turn a sequence (list) of elements into a multiset and vice-versa. These operations turn out to play an important role in connecting multisets and distributions. That's why we use this first section to look closer than we have done before. We shall see the role that permutations play. The next section introduces a 'zip' operation for multisets, in terms of accumulation and arrangement.

In the previous two chapters we have introduced the accumulation function $acc\colon \mathcal{L}(X) \to \mathcal{N}(X)$ that turns a list into a multiset by counting the occurrences of each element in the list, see (1.31). For instance, $acc(a, b, a, a) = 3|a\rangle + 1|b\rangle$. Clearly, *acc* is a surjective function. Accumulation forms a map of monads, see Exercise 1.9.7, which means that it is natural and commutes with the unit and flatten maps (of $\mathcal{L}$ and $\mathcal{M}$). In this chapter we shall use the accumlation map restricted to a particular 'size', via a number $K \in \mathbb{N}$ and via a restriction $acc[K]\colon X^K \to \mathcal{N}[K](X)$ to lists and multisets of size K. The parameter K in $acc[K]$ is often omitted if it is clear from the context.

In the other direction, one can go from multisets to lists/sequences via what we have called arrangement (2.10). This is not a function but channel, which

assigns equal probability to each sequence. Here we shall also use arrangement for a specific size, as a channel $arr[K]\colon \mathcal{N}[K](X) \nrightarrow X^K$. We recall:

$$arr[K](\varphi) = \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{\mathbf{(}\varphi\mathbf{)}} \big|\vec{x}\big\rangle \quad \text{where} \quad \mathbf{(}\varphi\mathbf{)} = \frac{\|\varphi\|!}{\varphi!} = \frac{K!}{\prod_x \varphi(x)!}.$$

We recall also that $\mathbf{(}\varphi\mathbf{)}$ is called the multiset coefficient of φ. It is the number of sequences that accumulate to φ, see Proposition 1.6.3.

One has $acc \circ arr = id$, see (2.11). Composition in the other direction does not give an identity but a (uniform) distribution of permutations:

$$(arr \circ acc)(\vec{x}) = \sum_{\vec{y}\in acc^{-1}(acc(\vec{x}))} \frac{1}{\mathbf{(}acc(\vec{x})\mathbf{)}} \big|\vec{y}\big\rangle = \sum_{\vec{y} \text{ is a permutation of } \vec{x}} \frac{1}{K!} \big|\vec{y}\big\rangle. \tag{3.5}$$

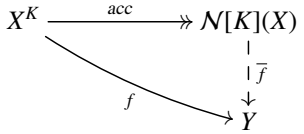
The vectors $\vec{y}$ take the multiplicities of elements in $\vec{x}$ into account, which leads to the factor $\frac{1}{\mathbf{(}acc(\vec{x})\mathbf{)}}$.

Permutations are an important part of the story of accumulation and arrangement. This is very explicit in the following result. Categorically it can be described in terms of a so-called coequaliser, but here we prefer a more concrete description. The axiomatic approach of [80] is based on this coequaliser.

Proposition 3.1.1. *For a number $K \in \mathbb{N}$, let $f\colon X^K \to Y$ be a function which is* stable under permutation*: for each permutation $\pi\colon \{1,\dots,K\} \xrightarrow{\cong} \{1,\dots,K\}$ one has*

$$f(x_1,\dots,x_K) = f(x_{\pi(1)},\dots,x_{\pi(K)}), \quad \text{for all sequences } (x_1,\dots,x_K) \in X^K.$$

Then there is a unique function $\overline{f}\colon \mathcal{N}[K](X) \to Y$ with $\overline{f} \circ acc = f$. In a diagram, we write this unique existence as a dashed arrow:



The double head $\twoheadrightarrow$ for acc is used to emphasise that it is a surjective function.

This result will be used both as a *definition principle* and as a *proof principle*. The existence part can be used to define a function $\mathcal{N}[K] \to Y$ by specifying a function $X^K \to Y$ that is stable under permutation. The uniqueness part yields a proof principle: if two functions $g_1, g_2\colon \mathcal{N}[K](X) \to Y$ satisfy $g_1 \circ acc = g_2 \circ acc$, then $g_1 = g_2$. This is quite powerful, as we shall see. Notice that acc is stable under permutation itself, see also Exercise 1.6.4.

Proof. Take $\varphi = \sum_{1 \leq i \leq \ell} n_i |x_i\rangle \in \mathcal{N}[K](X)$. Then we can define $\overline{f}(\varphi)$ via any arrangement, such as:

$$\overline{f}(\varphi) \coloneqq f(\underbrace{x_1, \ldots, x_1}_{n_1 \text{ times}}, \ldots, \underbrace{x_\ell, \ldots, x_\ell}_{n_\ell \text{ times}}).$$

Since f is stable under permutation, any other arrangement of the elements x_i gives the same outcome, that is, $\overline{f}(\varphi) = f(\vec{x})$ for each $\vec{x} \in X^K$ with $acc(\vec{x}) = \varphi$.

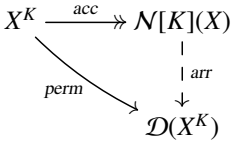
With this definition, $\overline{f} \circ acc = f$ holds, since for a sequence $\vec{x} = (x_1, \ldots, x_K) \in X^K$, say with $acc(\vec{x}) = \varphi$,

$$f(x_1, \ldots, x_K) = \overline{f}(\varphi) = \overline{f}(acc(\vec{x})) = (\overline{f} \circ acc)(x_1, \ldots, x_K).$$

For uniqueness, suppose $g \colon \mathcal{N}[K](X) \to Y$ also satisfies $g \circ acc = f$. Then $g = \overline{f}$, by the following argument. Take $\varphi \in \mathcal{N}[K](X)$; since acc is surjective, we can find $\vec{x} \in X^K$ with $acc(\vec{x}) = \varphi$. Then:

$$g(\varphi) = g(acc(\vec{x})) = f(\vec{x}) = \overline{f}(\varphi). \qquad \blacksquare$$

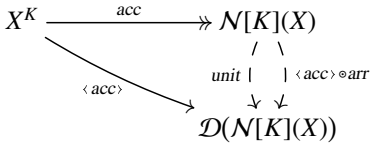
Example 3.1.2. We illustrate the use of Proposition 3.1.1 to re-define the arrangement map and to prove one of its basic properties. Assume for a moment that we do not already now about arrangement, only about accumulation. Now consider the following situation:



where $perm(\vec{x}) \coloneqq \sum_{\vec{y} \text{ is a permutation of } \vec{x}} \frac{1}{K!} |\vec{y}\rangle$.

By construction, the function *perm* is stable under permutation. Hence the function *acc* exists, by using Proposition 3.1.1 as a definition principle. As a result, $perm = arr \circ acc$.

Next we use Proposition 3.1.1 as proof principle. Our aim is to re-prove the equality $\langle acc \rangle \circ arr = unit \colon \mathcal{N}[K](X) \to \mathcal{D}(\mathcal{N}[K](X))$, which we already know from (2.11). Via our new proof principle such an equality can be obtained when both sides of the equation fit in a situation where there is a unique solution. This situation is given below.



We show that both dashed arrows fit. The first equation below holds by construction of *arr*, in the above triangle.

$$\begin{aligned}
\Big((\langle acc\rangle \circledcirc arr) \circ acc\Big)(\vec{x}) &= (\langle acc\rangle \circledcirc perm)(\vec{x}) \\
&= \mathcal{D}(acc)\left(\sum_{\vec{y} \text{ is a permutation of } \vec{x}} \frac{1}{K!}\big|\vec{y}\rangle\right) \\
&= \sum_{\vec{y} \text{ is a permutation of } \vec{x}} \frac{1}{K!}\big|\, acc(\vec{y})\,\rangle \\
&= \sum_{\vec{y} \text{ is a permutation of } \vec{x}} \frac{1}{K!}\big|\, acc(\vec{x})\,\rangle \\
&= 1\big|\, acc(\vec{x})\,\rangle \\
&= \langle acc\rangle(\vec{x}) \\
&= (unit \circ acc)(\vec{x}).
\end{aligned}$$

The fact that the composite $arr \circledcirc acc$ produces all permutation is used in the following result. Recall that the 'big' tensor $\bigotimes\colon \mathcal{D}(X)^K \to \mathcal{D}(X^K)$ is defined by $\bigotimes(\omega_1, \ldots, \omega_K) = \omega_1 \otimes \cdots \otimes \omega_K$, see 2.21.

Proposition 3.1.3. *The composite* arr $\circledcirc$ acc *commutes with tensors, as in the following diagram of channels.*

$$\begin{array}{ccccc}
\mathcal{D}(X)^K & \xrightarrow{acc} & \mathcal{N}[K](\mathcal{D}(X)) & \xrightarrow{arr} & \mathcal{D}(X)^K \\
\big\downarrow{\scriptstyle \bigotimes} & & & & \big\downarrow{\scriptstyle \bigotimes} \\
X^K & \xrightarrow{acc} & \mathcal{N}[K](X) & \xrightarrow{arr} & X^K
\end{array}$$

Proof. For distributions $\omega_1, \ldots, \omega_K \in \mathcal{D}(X)$ and elements $x_1, \ldots, x_K \in X$ we have:

$$\begin{aligned}
(\textstyle\bigotimes \circledcirc arr \circledcirc acc)(\vec{\omega})(\vec{x}) &= \sum_{\vec{\rho} \in acc^{-1}(acc(\vec{\omega}))} \frac{1}{\big(acc(\vec{\omega})\big)} \cdot (\rho_1 \otimes \cdots \otimes \rho_K)(\vec{x}) \\
&= \sum_{\vec{\rho} \text{ is a permutation of } \vec{\omega}} \frac{1}{K!} \cdot (\rho_1 \otimes \cdots \otimes \rho_K)(\vec{x}) \\
&= \sum_{\vec{y} \text{ is a permutation of } \vec{x}} \frac{1}{K!} \cdot (\omega_1 \otimes \cdots \otimes \omega_K)(\vec{y}) \\
&= \sum_{\vec{y} \in acc^{-1}(acc(\vec{x}))} \frac{1}{\big(acc(\vec{x})\big)} \cdot (\omega_1 \otimes \cdots \otimes \omega_K)(\vec{y}) \\
&= (arr \circledcirc acc \circledcirc \textstyle\bigotimes)(\vec{\omega})(\vec{x}).
\end{aligned}$$

❑

Exercises

3.1.1 Show that the permutation channel *perm* from Example 3.1.2 is an idempotent, *i.e.* satisfies $perm \circ perm = perm$. Prove this both concretely, via the definition of *perm*, and abstractly, via *acc* and *arr*.

3.1.2 Consider Proposition 3.1.3 for $X = \{a, b\}$ and $K = 4$. Check that:

$$\begin{aligned}
&(\bigotimes \circ arr \circ acc)(\omega_1, \omega_1, \omega_2, \omega_2)(a, b, b, b) \\
&= \tfrac{1}{2} \cdot \omega_1(a) \cdot \omega_1(b) \cdot \omega_2(b) \cdot \omega_2(b) + \tfrac{1}{2} \cdot \omega_1(b) \cdot \omega_1(b) \cdot \omega_2(a) \cdot \omega_2(b) \\
&= (arr \circ acc \circ \bigotimes)(\omega_1, \omega_1, \omega_2, \omega_2)(a, b, b, b).
\end{aligned}$$

3.2 Zipping multisets

For two multisets $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](Y)$ we can form their tensor product $\varphi \otimes \psi \in \mathcal{N}[K \cdot L](X \times Y)$. The fact that it has size $K \cdot L$ is shown in the first lines of the proof of Proposition 2.4.7. For sequences there is the familiar *zip* combination map $X^K \times Y^K \to (X \times Y)^K$ that does maintain size, see Exercise 1.1.7. Interestingly, there is also a zip-like operation for combining two multisets of the same size. It makes systematic use of accumulation and arrangement. This will be described next, based on [78].

Zipping two lists of the same length is a standard operation in (functional) programming. It produces a new list, consisting of pairs of elements from the two lists. We have seen it for products in Exercise 1.1.7, as an isomorphism $zip[K] \colon X^K \times Y^K \xrightarrow{\cong} (X \times Y)^K$. Our aim below is to describe a similar function, called multizip and written as *mzip*, for (natural) multisets of size K. It is not immediately obvious how *mzip* should work, since there are many ways to combine elements from two multisets. Analogously to the arrange channel $arr[K] \colon \mathcal{N}[K](X) \nrightarrow X^K$ we shall describe *mzip* on multisets as a (uniform) *channel*:

$$\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \xrightarrow{\quad mzip[K] \quad\circ\!\!\to} \mathcal{N}[K](X \times Y).$$

So let multisets $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[K](Y)$ be given. For all sequences $\vec{x} \in X^K$ and $\vec{y} \in Y^K$ with $acc(\vec{x}) = \varphi$ and $acc(\vec{y}) = \psi$ we can form their ordinary zip, as $zip(\vec{x}, \vec{y}) \in (X \times Y)^K$. Accumulating the pairs in this zip gives a multiset over $X \times Y$. Thus we define:

$$mzip(\varphi, \psi) \coloneqq \sum_{\vec{x} \in acc^{-1}(\varphi)} \sum_{\vec{y} \in acc^{-1}(\psi)} \frac{1}{(\varphi) \cdot (\psi)} \, \big| acc(zip(\vec{x}, \vec{y})) \big\rangle. \tag{3.6}$$

Diagrammatically we may describe *mzip* as the following composite.

$$\begin{array}{ccc} \mathcal{N}[K](X) \times \mathcal{N}[K](Y) & \xrightarrow{\mathit{arr} \otimes \mathit{arr}} & \mathcal{D}(X^K \times Y^K) \\ & & \cong \big\downarrow \mathcal{D}(\mathit{zip}) \\ & & \mathcal{D}((X \times Y)^K) \xrightarrow[\mathcal{D}(\mathit{acc})]{} \mathcal{D}(\mathcal{N}[K](X \times Y)) \end{array} \tag{3.7}$$

An illustration may help to see what happens here.

Example 3.2.1. Let's use two set $X = \{a, b\}$ and $Y = \{0, 1\}$ with two multisets of size three:

$$\varphi \;=\; 1|a\rangle + 2|b\rangle \qquad \text{and} \qquad \psi \;=\; 2|0\rangle + 1|1\rangle.$$

Then:

$$(\!|\varphi|\!) = \tbinom{3}{1,2} = 3 \qquad (\!|\psi|\!) = \tbinom{3}{2,1} = 3$$

The sequences in X^3 and Y^3 that accumulate to φ and ψ are:

$$\begin{cases} a, b, b \\ b, a, b \\ b, b, a \end{cases} \qquad \text{and} \qquad \begin{cases} 0, 0, 1 \\ 0, 1, 0 \\ 1, 0, 0 \end{cases}$$

Zipping them together gives the following nine sequences in $(X \times Y)^3$.

$$\begin{array}{lll} (a,0),(b,0),(b,1) & (b,0),(a,0),(b,1) & (b,0),(b,0),(a,1) \\ (a,0),(b,1),(b,0) & (b,0),(a,1),(b,0) & (b,0),(b,1),(a,0) \\ (a,1),(b,0),(b,0) & (b,1),(a,0),(b,0) & (b,1),(b,0),(a,0) \end{array}$$

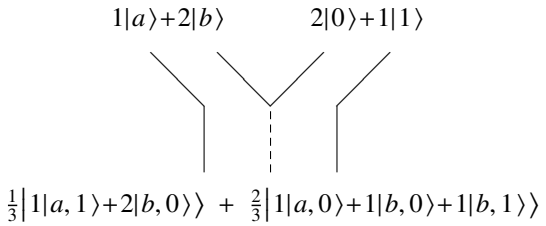
By applying the accumulation function *acc* to each of these we get multisets:

$$\begin{array}{ccc} 1|a,0\rangle+1|b,0\rangle+1|b,1\rangle & 1|b,0\rangle+1|a,0\rangle+1|b,1\rangle & 2|b,0\rangle+1|a,1\rangle \\ 1|a,0\rangle+1|b,1\rangle+1|b,0\rangle & 2|b,0\rangle+1|a,1\rangle & 1|b,0\rangle+1|b,1\rangle+1|a,0\rangle \\ 1|a,1\rangle+2|b,0\rangle & 1|b,1\rangle+1|a,0\rangle+1|b,0\rangle & 1|b,1\rangle+1|b,0\rangle+1|a,0\rangle \end{array}$$

We see that are only two different multisets involved. Counting them and multiplying with $\frac{1}{(\!|\varphi|\!)(\!|\psi|\!)} = \frac{1}{9}$ gives:

$$\begin{aligned} &\mathit{mzip}[3]\big(1|a\rangle+2|b\rangle,\; 2|0\rangle+1|1\rangle\big) \\ &= \tfrac{1}{3}\big|\,1|a,1\rangle+2|b,0\rangle\,\big\rangle \;+\; \tfrac{2}{3}\big|\,1|a,0\rangle+1|b,0\rangle+1|b,1\rangle\,\big\rangle. \end{aligned}$$

This shows that calculating *mzip* is laborious. But it is quite mechanical and easy to implement. The picture below suggests to look at *mzip* as a funnel with two input pipes in which multiple elements from both sides can be combined into a probabilistic mixture.



The *mzip* operation satisfies several useful properties.

Proposition 3.2.2. *Consider* $\mathrm{mzip}\colon \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \nrightarrow \mathcal{N}[K](X \times Y)$, *either in equational formulation* (3.6) *or in diagrammatic form* (3.7).

1 *The* mzip *map is natural in* X *and* Y: *for functions* $f\colon X \to U$ *and* $g\colon Y \to V$ *the following diagram commutes.*

$$\begin{array}{ccc} \mathcal{N}[K](X) \times \mathcal{N}[K](Y) & \xrightarrow{mzip} & \mathcal{D}(\mathcal{N}[K](X \times Y)) \\ {\scriptstyle \mathcal{N}[K](f)\times\mathcal{N}[K](g)}\downarrow & & \downarrow{\scriptstyle \mathcal{D}(\mathcal{N}[K](f\times g))} \\ \mathcal{N}[K](U) \times \mathcal{N}[K](V) & \xrightarrow{mzip} & \mathcal{D}(\mathcal{N}[K](U \times V) \end{array}$$

2 *For* $\varphi \in \mathcal{N}[K](X)$ *and* $y \in Y$,

$$\mathrm{mzip}(\varphi, K|y\rangle) \;=\; 1\big|\varphi \otimes 1|y\rangle\big\rangle.$$

And similarly in symmetric form.

3 *Multizip commutes with projections in the following sense.*

$$\begin{array}{ccccc} \mathcal{N}[K](X) & \xleftarrow{\pi_1} & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) & \xrightarrow{\pi_2} & \mathcal{N}[K](Y) \\ {\scriptstyle unit}\downarrow & & \downarrow{\scriptstyle mzip} & & \downarrow{\scriptstyle unit} \\ \mathcal{D}(\mathcal{N}[K](X)) & \xleftarrow{\mathcal{D}(\mathcal{N}[K](\pi_1))} & \mathcal{D}(\mathcal{N}[K](X \times Y)) & \xrightarrow{\mathcal{D}(\mathcal{N}[K](\pi_2))} & \mathcal{D}(\mathcal{N}[K](Y)) \end{array}$$

4 *Arrangement* arr *relates* zip *and* mzip *as in:*

$$\begin{array}{ccc} \mathcal{N}[K](X) \times \mathcal{N}[K](Y) & \xrightarrow{arr\otimes arr} & X^X \times Y^K \\ {\scriptstyle mzip}\downarrow & & \downarrow{\scriptstyle zip} \\ \mathcal{N}[K](X \times Y) & \xrightarrow{arr} & (X \times Y)^K \end{array}$$

5 mzip *is associative, as given by:*

$$\begin{array}{ccc} \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \times \mathcal{N}[K](Z) & \xrightarrow{id\otimes mzip} & \mathcal{N}[K](X) \times \mathcal{N}[K](Y \times Y) \\ {\scriptstyle mzip\otimes id}\downarrow & & \downarrow{\scriptstyle mzip} \\ \mathcal{N}[K](X \times Y) \times \mathcal{N}[K](Z) & \xrightarrow{mzip} & \mathcal{N}[K](X \times Y \times Z) \end{array}$$

Here we take associativity of $\times$ *for granted.*

Proof. 1 Easy, via the diagrammatic formulation (3.7), using naturality of *arr* (Exercise 2.2.7), of *zip* (Exercise 1.9.4), and of *acc* (Exercise 1.6.6).

2 Since:

$$\begin{aligned} mzip(\varphi, K|y\rangle) &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|acc(zip(\vec{x}, \langle y, \ldots, y\rangle))\big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|acc(\overrightarrow{x_i, y})\big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|acc(\vec{x}) \otimes 1|y\rangle\big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|\varphi \otimes 1|y\rangle\big\rangle \;=\; 1\big|\varphi \otimes 1|y\rangle\big\rangle. \end{aligned}$$

3 By naturality of *acc* and *zip*:

$$\begin{aligned} &\mathcal{D}(\mathcal{N}[K](\pi_1))\Big(mzip(\varphi, \psi)\Big) \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\varphi)\cdot(\psi)} \big|\mathcal{N}[K](\pi_1)(acc(zip(\vec{x}, \vec{y})))\big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\varphi)\cdot(\psi)} \big|acc((\pi_1)^K(zip(\vec{x}, \vec{y})))\big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\varphi)\cdot(\psi)} \big|acc(\vec{x})\big\rangle \;=\; \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|\varphi\big\rangle \;=\; 1\big|\varphi\big\rangle. \end{aligned}$$

4 We have for $\varphi \in \mathcal{N}[K](X)$, $\psi \in \mathcal{N}[K](Y)$,

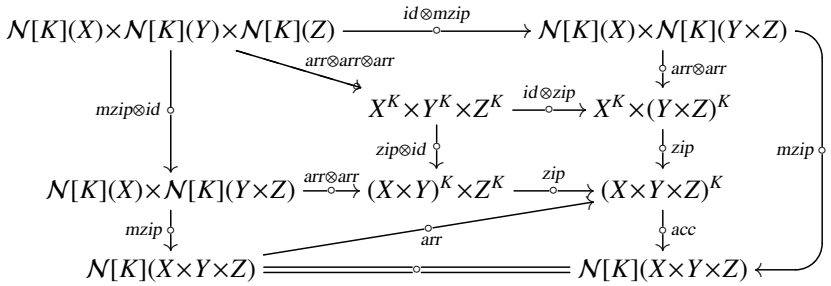
$$\begin{aligned} (arr \circ mzip)(\varphi, \psi) &= \sum_{\vec{z}\in(X\times Y)^K} \left(\sum_{\chi\in\mathcal{N}[K](X\times Y)} arr(\chi)(\vec{z}) \cdot mzip(\varphi, \psi)(\chi) \right) \big|\vec{z}\big\rangle \\ &= \sum_{\vec{z}\in(X\times Y)^K} \frac{1}{(acc(\vec{z}))} \cdot mzip(\varphi, \psi)(acc(\vec{z}))\big|\vec{z}\big\rangle \\ &= \sum_{\vec{z}\in(X\times Y)^K} \frac{1}{(acc(\vec{z}))} \cdot \left(\sum_{\vec{x}\in acc^{-1}(\varphi), \vec{y}\in acc^{-1}(\psi), acc(zip(\vec{x},\vec{y}))=acc(\vec{z})} \frac{1}{(\varphi)} \cdot \frac{1}{(\psi)} \right) \big|\vec{z}\big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\varphi)} \cdot \frac{1}{(\psi)} \cdot \sum_{\vec{z}\in acc^{-1}(acc(zip(\vec{x},\vec{y})))} \frac{1}{(acc(zip(\vec{x}, \vec{y})))} \big|\vec{z}\big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\varphi)} \cdot \frac{1}{(\psi)} \big|zip(\vec{x}, \vec{y})\big\rangle \end{aligned}$$

since permuting a zip of permuted inputs does not add anything

$$= (\langle zip\rangle \circ (arr \otimes arr))(\varphi, \psi).$$

5 Consider the following diagram chase, obtained by unpacking the (diagram-

matic) definition of *mzip* on the right-hand side.



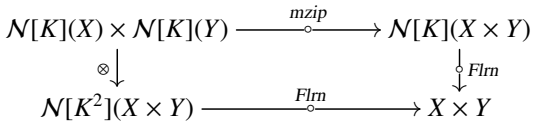
We can see that the outer diagram commutes by going through the internal subdiagrams. In the middle we use associativity of the (ordinary) *zip* function, formulated in terms of (deterministic) channels. Three of the (other) internal subdiagrams commute by item (4). The *acc*-*arr* triangle at the bottom commutes by (2.11). ❑

The following result deserves a separate status. It tells that what we learn from a multiset zip is the same as what we learn from a parallel product (of multisets).

Theorem 3.2.3. *Multiset zip and frequentist learning interact well, namely as:*

$$Flrn \gg\!= mzip(\varphi, \psi) \;=\; Flrn(\varphi \otimes \psi).$$

Equivalently, in diagrammatic form:



Proof. Let multisets $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[K](Y)$ be given and let $a \in X$ and $b \in Y$ be arbitrary elements. We need to show that the probability:

$$\Big(Flrn \gg\!= mzip(\varphi, \psi)\Big)(a, b) \;=\; \sum_{\vec{x} \in acc^{-1}(\varphi)} \; \sum_{\vec{y} \in acc^{-1}(\psi)} \frac{acc(zip(\vec{x}, \vec{y}))(a, b)}{K \cdot \mathbf{(}\varphi\mathbf{)} \cdot \mathbf{(}\psi\mathbf{)}}$$

is the same as the probability:

$$Flrn(\varphi \otimes \psi)(a, b) \;=\; \frac{\varphi(a) \cdot \psi(a)}{K \cdot K}.$$

We reason informally, as follows. For arbitrary $\vec{x} \in acc^{-1}(\varphi)$ and $\vec{y} \in acc^{-1}(\psi)$ we need to find the fraction of occurrences (a, b) in $zip(\vec{x}, \vec{y})$. The fraction of occurrences of a in $\vec{x}$ is $\frac{\varphi(a)}{K} = Flrn(\varphi)(a)$, and the fraction of occurrences of b

in $\vec{y}$ is $\frac{\psi(b)}{K} = \mathit{Flrn}(\psi)(b)$. Hence the fraction of occurrences of (a, b) in $\mathit{zip}(\vec{x}, \vec{y})$ is $\mathit{Flrn}(\varphi)(a) \cdot \mathit{Flrn}(\psi)(b) = \mathit{Flrn}(\varphi \otimes \psi)(a, b)$. ❑

Once we have seen the definition of *mzip*, via 'deconstruction' of multisets into lists, a *zip* operation on lists, and 'reconstruction' to a multiset result, we can try to apply this approach more widely. For instance, instead of using a *zip* on lists we can simply concatenate (++) the lists — assuming they contain elements from the same set. This yields, like in (3.7), a composite channel:

$$\begin{array}{ccc} \mathcal{N}[K](X) \times \mathcal{N}[L](X) & \xrightarrow{\mathit{arr} \otimes \mathit{arr}} & \mathcal{D}(X^K \times X^L) \\ & & \cong \big\downarrow \mathcal{D}(+\!+) \\ & & \mathcal{D}(X^{K+L}) \xrightarrow[\mathcal{D}(\mathit{acc})]{} \mathcal{D}(\mathcal{N}[K+L](X)) \end{array}$$

It is easy to see that this yields addition of multisets, as a deterministic channel.

We don't get the tensor $\otimes$ of multisets in this way, because there is no tensor of lists, see Remark 2.4.3.

We conclude with several useful observations about accumulation in two dimensions.

Lemma 3.2.4. *Fix a number $K \in \mathbb{N}$ and a set X.*

1 We can mix K-ary and 2-ary accumulation in the following way.

$$\begin{array}{ccccc} X^K \times X^K & \xrightarrow{\mathit{acc}[K] \times \mathit{acc}[K]} & \mathcal{N}[K](X) \times \mathcal{N}[K](X) & \xrightarrow{+} & \\ \mathit{zip} \big\downarrow \cong & & & & \mathcal{N}[2K](X) \\ (X \times X)^K & \xrightarrow{\mathit{acc}[2]^K} & \mathcal{N}[2](X)^K & \xrightarrow{+} & \end{array}$$

2 When we generalise from 2 to $L \geq 2$ we get:

$$\begin{array}{ccccc} (X^K)^L & \xrightarrow{\mathit{acc}[K]^L} & \mathcal{N}[K](X)^L & \xrightarrow{+} & \\ \mathit{zip}_L \big\downarrow \cong & & & & \mathcal{N}[L \cdot K](X) \\ (X^L)^K & \xrightarrow{\mathit{acc}[L]^K} & \mathcal{N}[L](X)^K & \xrightarrow{+} & \end{array}$$

Proof. 1 Commutation of the diagram is 'obvious', so we provide only an

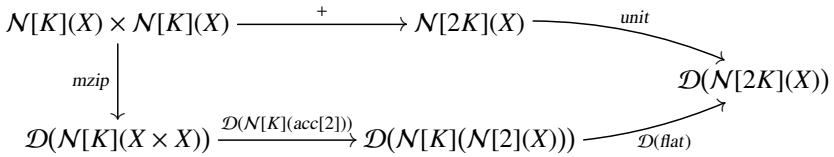
exemplary proof. The two paths in the diagram yield the same outcomes in:

$$\begin{aligned}
&(+ \circ (acc[K] \times acc[K]))([a,b,c,a,c],[b,b,a,a,c]) \\
&\quad = (2|a\rangle + 1|b\rangle + 2|c\rangle) + (2|a\rangle + 2|b\rangle + 1|c\rangle) \\
&\quad = 4|a\rangle + 3|b\rangle + 3|c\rangle. \\
&(+ \circ acc[2]^K \circ zip)([a,b,c,a,c],[b,b,a,a,c]) \\
&\quad = (+ \circ acc[2]^K)([(a,b),(b,b),(c,a),(a,a),(c,c)]) \\
&\quad = +([1|a\rangle + 1|b\rangle, 2|b\rangle, 1|a\rangle + 1|c\rangle, 2|a\rangle, 2|c\rangle]) \\
&\quad = 4|a\rangle + 3|b\rangle + 3|c\rangle.
\end{aligned}$$

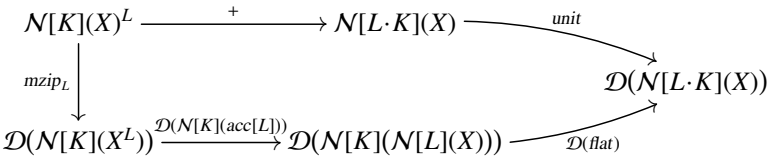
2 Similarly. ❑

We now transfer these results to multizip.

Proposition 3.2.5. *In binary form one has:*



More generally, we have for $L \geq 2$*,*



The map mzip_L *is the L-ary multizip, obtained via:*

$$mzip_2 \coloneqq mzip \quad \text{and} \quad mzip_{L+1} \coloneqq mzip \circ (mzip_L \otimes id)$$

Via the associativity of Proposition 3.2.2 (5) *the actual arrangement of these multiple multizips does not matter.*

Proof. We only do the binary case, via an equational proof:

$$
\begin{aligned}
&\mathcal{D}(\mathit{flat}) \circ \mathcal{D}(\mathcal{N}[K](\mathit{acc}[2])) \circ \mathit{mzip} \\
\overset{(3.7)}{=}\ & \mathcal{D}(\mathit{flat}) \circ \mathcal{D}(\mathcal{N}[K](\mathit{acc}[2])) \circ \mathcal{D}(\mathit{acc}[K]) \circ \mathcal{D}(\mathit{zip}) \circ (\mathit{arr} \otimes \mathit{arr}) \\
=\ & \mathcal{D}(\mathit{flat}) \circ \mathcal{D}(\mathit{acc}[K]) \circ \mathcal{D}(\mathit{acc}[2]^K) \circ \mathcal{D}(\mathit{zip}) \circ (\mathit{arr} \otimes \mathit{arr}) \\
& \qquad \text{by naturality of } \mathit{acc}[K] \colon X^K \to \mathcal{N}[K](X) \\
=\ & \mathcal{D}(+) \circ \mathcal{D}(\mathit{acc}[2]^K) \circ \mathcal{D}(\mathit{zip}) \circ (\mathit{arr} \otimes \mathit{arr}) \qquad \text{by Exercise 1.6.7} \\
=\ & \mathcal{D}(+) \circ \mathcal{D}(\mathit{acc}[K] \times \mathit{acc}[K]) \circ (\mathit{arr} \otimes \mathit{arr}) \qquad \text{by Lemma 3.2.4 (1)} \\
=\ & \mathcal{D}(+) \circ ((\mathcal{D}(\mathit{acc}[K]) \circ \mathit{arr}) \otimes (\mathcal{D}(\mathit{acc}[K]) \circ \mathit{arr})) \\
\overset{(2.11)}{=}\ & \mathcal{D}(+) \circ (\mathit{unit} \otimes \mathit{unit}) \\
=\ & \mathcal{D}(+) \circ \mathit{unit} \\
=\ & \mathit{unit} \circ +.
\end{aligned}
$$

❑

Exercises

3.2.1 Show that:

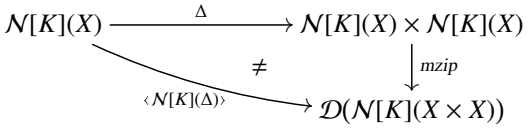
$$
\begin{aligned}
&\mathit{mzip}[4]\big(1|a\rangle + 2|b\rangle + 1|c\rangle,\ 3|0\rangle + 1|1\rangle\big) \\
&= \tfrac{1}{4}\big|\, 1|a, 1\rangle + 2|b, 0\rangle + 1|c, 0\rangle\big\rangle \\
&\qquad + \tfrac{1}{2}\big|\, 1|a, 0\rangle + 1|b, 0\rangle + 1|b, 1\rangle + 1|c, 0\rangle\big\rangle \\
&\qquad + \tfrac{1}{4}\big|\, 1|a, 0\rangle + 2|b, 0\rangle + 1|c, 1\rangle\big\rangle.
\end{aligned}
$$

3.2.2 Show in the context of the previous exercise that:

$$
\begin{aligned}
&\mathit{Flrn} \gg\!= \mathit{mzip}[4]\big(1|a\rangle + 2|b\rangle + 1|c\rangle,\ 3|0\rangle + 1|1\rangle\big) \\
&= \tfrac{3}{16}|a, 0\rangle + \tfrac{1}{16}|a, 1\rangle + \tfrac{3}{8}|b, 0\rangle + \tfrac{1}{8}|b, 1\rangle + \tfrac{3}{16}|c, 0\rangle + \tfrac{1}{16}|c, 1\rangle.
\end{aligned}
$$

Compute also: $\mathit{Flrn}(1|a\rangle + 2|b\rangle + 1|c\rangle) \otimes \mathit{Flrn}(3|0\rangle + 1|1\rangle)$, and remember Theorem 3.2.3.

3.2.3 1 Check that *mzip* does *not* commute with diagonals, in the sense that the following triangle does *not* commute.



Hint: Take for instance $1|a\rangle + 1|b\rangle$.

2 Check that *mzip* and *zip* do *not* commute with accumulation, as in:

$$\begin{array}{ccc} X^K \times Y^K & \xrightarrow{\text{acc} \times \text{acc}} & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \\ {\scriptstyle zip}\downarrow{\scriptstyle \cong} & \neq & \downarrow{\scriptstyle mzip} \\ (X \times Y)^K & \xrightarrow{\langle \text{acc} \rangle} & \mathcal{D}(\mathcal{N}[K](X \times Y)) \end{array}$$

Hint: Take sequences $[a, b, b]$, $[0, 0, 1]$ and re-use Example 3.2.1.

3.3 The multinomial channel

Multinomial distributions appear when one draws multiple coloured balls from an urn, with replacement, as described briefly in Example 2.1.1 (2). These distributions assign a probability to such a draw, where the draws themselves will be represented as multisets, over colours. We shall standardly describe multinomial distributions in multivariate form, for multiple colours. The binomial distribution is then a special case, for two colours only, see Example 2.1.1 (2).

A multinomial draw is a draw 'with replacement', so that the draw of multiple, say K, balls can be understood as K consecutive draws from the same urn. Such draws may also be understood as samples from the distribution, of size K. This addititive nature of multinomials is formalised as Exercise 3.3.8 below. In different form it occurs in Section 3.5.

When one thinks in terms of drawing from an urn with replacement, the urn itself never changes. For that reason, the urn is most conveniently — and abstractly — represented as a distribution, rather than as a multiset. When the urn is written as $\omega \in \mathcal{D}(X)$, the underlying set/space X is the set of types (think: colours) of objects in the urn. To sum up, multinomial distributions are described as a function of the following type.

$$\mathcal{D}(X) \xrightarrow{mn[K]} \mathcal{D}\big(\mathcal{N}[K](X)\big). \tag{3.8}$$

The number $K \in \mathbb{N}$ represents the number of objects that is drawn. The distribution $mn[K](\omega)$ assigns a probability to a K-object draw $\varphi \in \mathcal{N}[K](X)$. There is no bound on K, since the idea is that drawn objects are replaced.

Clearly, the above function (3.8) forms a channel $mn[K] \colon \mathcal{D}(X) \nrightarrow \mathcal{N}[K](X)$. In this section we collect some basic properties of these channels. They are interesting in themselves, since they capture basic relationships between multisets and distributions, but they will also be useful in the sequel.

For convenience we repeat the definition of the multinomial channel (3.8) from Example 2.1.1 (2) For a set X and a natural number $K \in \mathbb{N}$ we define the

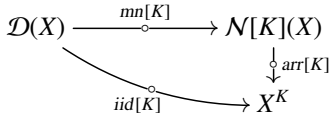
multinomial channel (3.8) on $\omega \in \mathcal{D}(X)$ as:

$$mn[K](\omega) \coloneqq \sum_{\varphi \in \mathcal{N}[K](X)} \binom{}{\varphi} \cdot \prod_x \omega(x)^{\varphi(x)} \,|\varphi\rangle, \tag{3.9}$$

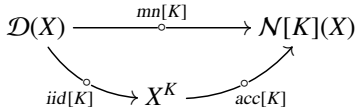
where, recall, $\binom{}{\varphi}$ is the multinomial coefficient $\frac{K!}{\prod_i \varphi(x_i)!}$, see Definition 1.5.1 (5). The Multinomial Theorem (1.26) ensures that the probabilities in $mn[K](\omega)$ add up to one.

The next result describes the fundamental relationships between multinomial distributions, accumulation/arrangement (*acc* / *arr*) and indedependent and identical distributions (*iid*).

Theorem 3.3.1. *1 Arrangement after a multinomial yields the indedependent and identical version of the original distribution:*



2 Multinomial channels can be described as composite:



where $iid(\omega) = \omega^K = \omega \otimes \cdots \otimes \omega \in \mathcal{D}(X^K)$.

Proof. 1 For $\omega \in \mathcal{D}(X)$ and $\vec{x} = (x_1, \ldots, x_K) \in X^K$,

$$\begin{aligned}
(arr \circ mn[K])(\omega)(\vec{x}) &= \sum_{\varphi \in \mathcal{N}[K](X)} arr(\varphi)(\vec{x}) \cdot mn[K](\omega)(\varphi) \\
&= \frac{1}{\binom{}{acc(\vec{x})}} \cdot mn[K](\omega)(acc(\vec{x})) \\
&= \frac{1}{\binom{}{acc(\vec{x})}} \cdot \binom{}{acc(\vec{x})} \cdot \prod_y \omega(y)^{acc(\vec{x})(y)} \\
&= \prod_i \omega(x_i) \;=\; \omega^K(\vec{x}) \;=\; iid(\omega)(\vec{x}).
\end{aligned}$$

2 A direct consequence of the previous item, using that $acc \circ arr = unit$, see (2.11) in Example 2.2.3.

This theorem has non-trivial consequences, such as naturality of multinomial channels and commutation with tensors.

Corollary 3.3.2. *Multinomial channels are natural: for each function $f\colon X \to Y$ the following diagram commutes.*

$$\begin{array}{ccc} \mathcal{D}(X) & \xrightarrow{mn[K]} & \mathcal{D}(\mathcal{N}[K](X)) \\ {\scriptstyle \mathcal{D}(f)}\downarrow & & \downarrow{\scriptstyle \mathcal{D}(\mathcal{N}[K](f))} \\ \mathcal{D}(X) & \xrightarrow{mn[K]} & \mathcal{D}(\mathcal{N}[K](X)) \end{array}$$

Proof. By Theorem 3.3.1 (2) and naturality of *acc* and of *iid*, as expressed by the diagram on the left in Lemma 2.4.8 (1).

$$\begin{aligned} \mathcal{D}(\mathcal{N}[K](f)) \circ mn[K] &= \mathcal{D}(\mathcal{N}[K](f)) \circ \mathcal{D}(acc) \circ iid \\ &= \mathcal{D}(acc) \circ \mathcal{D}(f^K) \circ iid \\ &= \mathcal{D}(acc) \circ iid \circ \mathcal{D}(f) \\ &= mn[K] \circ \mathcal{D}(f). \end{aligned}$$ ❑

For the next result, recall the multizip operation *mzip* from Section 3.2. It may have looked a bit unusual at the time, but the next result demonstrates that it behaves quite well — as in other such results.

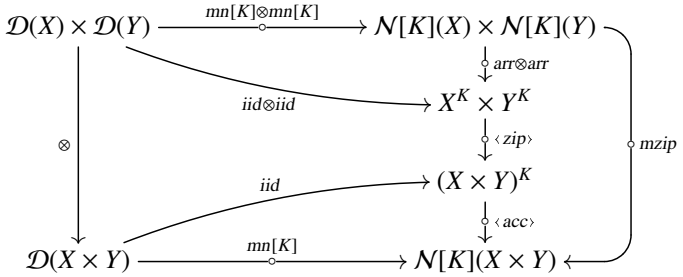
Corollary 3.3.3. *Multinomial channels commute with tensor and multizip:*

$$mzip \gg\!= \Big(mn[K](\omega) \otimes mn[K](\rho)\Big) = mn[K](\omega \otimes \rho).$$

Diagrammatically this amounts to:

$$\begin{array}{ccc} \mathcal{D}(X) \times \mathcal{D}(Y) & \xrightarrow{mn[K] \otimes mn[K]} & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \\ {\scriptstyle \otimes}\downarrow & & \downarrow{\scriptstyle mzip} \\ \mathcal{D}(X \times Y) & \xrightarrow{mn[K]} & \mathcal{N}[K](X \times Y) \end{array}$$

Proof. We give a proof via diagram-chasing. Commutation of the outer diagram below follows from the commuting subparts of the diagram, which arise by unfolding the definition of *mzip* in (3.7), below on the right.



The three subdiagrams, from top to bottom, commute by Theorem 3.3.1 (1), by Lemma 2.4.8 (2), and by Theorem 3.3.1 (2). ❑

Actually computing $mn[K](\omega \otimes \rho)(\chi)$ is very fast, but computing the equal expression:

$$\Big(mzip \gg\!= (mn[K](\omega) \otimes mn[K](\rho))\Big)(\chi)$$

is much much slower. The reason is that one has to sum over all pairs (φ, χ) that *mzip* to χ.

We move on to a next fundamental fact, namely that frequentist learning after a multinomial is the identity, in the theorem below. We first need an auxiliary result.

Lemma 3.3.4. *Fix a distribution $\omega \in \mathcal{D}(X)$ and a number K. For each $y \in X$,*

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \;=\; K \cdot \omega(y).$$

Proof. The equation holds for $K = 0$, since then $\varphi(y) = 0$. Hence we may assume $K > 0$. Then:

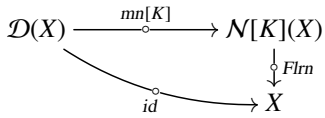
$$\begin{aligned}
&\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \\
&= \sum_{\varphi \in \mathcal{N}[K](X),\, \varphi(y) \neq 0} \varphi(y) \cdot \frac{K!}{\prod_x \varphi(x)!} \cdot \prod\nolimits_x \omega(x)^{\varphi(x)} \\
&= \sum_{\varphi \in \mathcal{N}[K](X),\, \varphi(y) \neq 0} \frac{K \cdot (K-1)!}{(\varphi(y)-1)! \cdot \prod_{x \neq y} \varphi(x)!} \cdot \omega(y) \cdot \omega(y)^{\varphi(y)-1} \cdot \prod_{x \neq y} \omega(x)^{\varphi(x)} \\
&= K \cdot \omega(y) \cdot \sum_{\varphi \in \mathcal{N}[K-1](X)} \frac{(K-1)!}{\prod_x \varphi(x)!} \cdot \prod\nolimits_x \omega(x)^{\varphi(x)} \\
&= K \cdot \omega(y) \cdot \sum_{\varphi \in \mathcal{N}[K-1](X)} mn[K-1](\omega)(\varphi) \\
&= K \cdot \omega(y).
\end{aligned}$$

❑

Theorem 3.3.5. *Frequentist learning from a multinomial gives the original distribution:*

$$Flrn \gg\!= mn[K](\omega) \;=\; \omega. \tag{3.10}$$

This means that the following diagram of channels commutes.



The identity function $\mathcal{D}(X) \to \mathcal{D}(X)$ is used as channel $\mathcal{D}(X) \nrightarrow X$.

This last result is important when one thinks of draws as samples from the distribution. It says that what we learn from samples is the same as what we

learn from the distribution itself. This is of course an elementary correctness property for sampling.

Proof. By Lemma 3.3.4:

$$\begin{aligned}(\mathit{Flrn} \circ \mathit{mn}[K])(\omega)(y) &= \sum_{\varphi \in \mathcal{N}[K](X)} \mathit{mn}[K](\omega)(\varphi) \cdot \mathit{Flrn}(\varphi)(y) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} \mathit{mn}[K](\omega)(\varphi) \cdot \frac{\varphi(y)}{\|\varphi\|} \\ &= \|\varphi\| \cdot \omega(y) \cdot \frac{1}{\|\varphi\|} \\ &= \omega(y).\end{aligned}$$

❑

Since a multinomial $\mathit{mn}[K](\omega)$ is a distribution we can use it as an abstract urn, not containing single balls, but containing multisets of balls (draws). Hence we can draw from $\mathit{mn}[K](\omega)$ as well, giving a distribution on draws of draws. We show that this can also be done with a single multinomial.

Theorem 3.3.6. *Multinomial channels compose, with a bit of help of the (fixed-size) flatten operation for multisets* (1.33)*, as in:*

$$\begin{array}{ccccc} \mathcal{D}(X) & \xrightarrow{\mathit{mn}[K]} & \mathcal{D}(\mathcal{M}[K](X)) & \xrightarrow{\mathit{mn}[L]} & \mathcal{D}\Big(\mathcal{M}[L](\mathcal{M}[K](X))\Big) \\ & \searrow^{\mathit{mn}[L\cdot K]} & & & \downarrow{\scriptstyle \mathcal{D}(\mathit{flat})} \\ & & & & \mathcal{D}\Big(\mathcal{M}[L\cdot K](X)\Big) \end{array}$$

Proof. For $\omega \in \mathcal{D}(X)$ and $\psi \in \mathcal{M}[L \cdot K](X)$,

$$\begin{aligned}
&\Big(\mathcal{D}(\mathit{flat}) \circ mn[L] \circ mn[K]\Big)(\omega)(\psi) \\
&= \sum_{\Psi \in \mathcal{M}[L](\mathcal{M}[K](X)),\, \mathit{flat}(\Psi) = \psi} mn[L](mn[K](\omega))(\Psi) \\
&= \sum_{\Psi \in \mathcal{M}[L](\mathcal{M}[K](X)),\, \mathit{flat}(\Psi) = \psi} \binom{}{\Psi} \cdot \prod_{\varphi \in \mathit{supp}(\Psi)} \Big(\binom{}{\varphi} \cdot \textstyle\prod_x \omega(x)^{\varphi(x)}\Big)^{\Psi(\varphi)} \\
&= \sum_{\Psi \in \mathcal{M}[L](\mathcal{M}[K](X)),\, \mathit{flat}(\Psi) = \psi} \binom{}{\Psi} \cdot \prod_{\varphi \in \mathit{supp}(\Psi)} \binom{}{\varphi}^{\Psi(\varphi)} \cdot \prod_x \prod_{\varphi \in \mathit{supp}(\Psi)} \omega(x)^{\varphi(x) \cdot \Psi(\varphi)} \\
&= \sum_{\Psi \in \mathcal{M}[L](\mathcal{M}[K](X)),\, \mathit{flat}(\Psi) = \psi} \binom{}{\Psi} \cdot \prod_{\varphi \in \mathit{supp}(\Psi)} \binom{}{\varphi}^{\Psi(\varphi)} \cdot \prod_x \omega(x)^{\sum_{\varphi \in \mathit{supp}(\Psi)} \varphi(x) \cdot \Psi(\varphi)} \\
&= \sum_{\Psi \in \mathcal{M}[L](\mathcal{M}[K](X)),\, \mathit{flat}(\Psi) = \psi} \binom{}{\Psi} \cdot \prod_{\varphi \in \mathit{supp}(\Psi)} \binom{}{\varphi}^{\Psi(\varphi)} \cdot \prod_x \omega(x)^{\mathit{flat}(\Psi)(x)} \\
&= \left(\sum_{\Psi \in \mathcal{M}[L](\mathcal{M}[K](X)),\, \mathit{flat}(\Psi) = \psi} \binom{}{\Psi} \cdot \prod_{\varphi \in \mathit{supp}(\Psi)} \binom{}{\varphi}^{\Psi(\varphi)} \right) \cdot \prod_x \omega(x)^{\psi(x)} \\
&= \binom{}{\psi} \cdot \textstyle\prod_x \omega(x)^{\psi(x)} \qquad \text{by Theorem 1.6.5} \\
&= mn[L \cdot K](\omega)(\psi).
\end{aligned}$$

❑

We mention another consequence of Lemma 3.3.4; it may be understood as an 'average' or 'mean' result for multinomials, see also Definition 4.1.3. We can describe it via a flatten map $\mathit{flat} \colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$, using inclusions $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ and $\mathcal{N}[K](X) \hookrightarrow \mathcal{M}(X)$.

Proposition 3.3.7. *Fix a distribution $\omega \in \mathcal{D}(X)$ and a number K. Each natural multiset $\varphi \in \mathcal{N}[K](X)$ can be regarded as an element of the set of all multisets $\mathcal{M}(X)$. Similarly, ω can be understood as an element of $\mathcal{M}(X)$. With these inclusions in mind one has:*

$$\mathit{flat}(mn[K](\omega)) = \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi = K \cdot \omega.$$

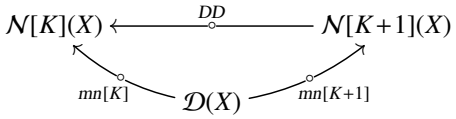
Proof. Since:

$$\begin{aligned}
\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi &= \sum_{x \in X} \left(\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x) \right) \big| x \big\rangle \\
&= \sum_{x \in X} K \cdot \omega(x) \, \big| x \big\rangle \qquad \text{by Lemma 3.3.4} \\
&= K \cdot \sum_{x \in X} \omega(x) \, \big| x \big\rangle \\
&= K \cdot \omega.
\end{aligned}$$

❑

Recall the draw-and-delete channel $DD\colon \mathcal{N}[K+1](X) \nrightarrow \mathcal{N}[K](X)$ from Definition 2.2.4 for drawing a single ball from a non-empty multiset. It commutes with multinomial channels.

Proposition 3.3.8. *The following triangles commute, for $K \in \mathbb{N}$.*



Proof. For $\omega \in \mathcal{D}(X)$ and $\varphi \in \mathcal{N}[K](X)$ we have:

$$\begin{aligned}
&(DD \mathbin{\circ} mn[K+1])(\omega)(\varphi) \\
&= \sum_{\psi \in \mathcal{N}[K+1](X)} mn[K+1](\omega)(\psi) \cdot DD[K](\psi)(\varphi) \\
&= \sum_{x \in X} mn[K+1](\omega)(\varphi + 1|x\rangle) \cdot \frac{\varphi(x)+1}{K+1} \\
&= \sum_{x \in X} \frac{(K+1)!}{\prod_y (\varphi + 1|x\rangle)(y)!} \cdot \prod\nolimits_y \omega(y)^{(\varphi+1|x\rangle)(y)} \cdot \frac{\varphi(x)+1}{K+1} \\
&= \sum_{x \in X} \frac{K!}{\prod_y \varphi(y)!} \cdot \prod\nolimits_y \omega(y)^{\varphi(y)} \cdot \omega(x) \\
&= \mathbf{(}\varphi\mathbf{)} \cdot \prod\nolimits_y \omega(y)^{\varphi(y)} \cdot \big(\textstyle\sum_x \omega(x)\big) \\
&= mn[K](\omega)(\varphi).
\end{aligned}$$

❑

The above triangles exist for each K. This means that the collection of channels $mn[K]$, indexed by $K \in \mathbb{N}$, forms a cone for the infinite chain of draw-and-delete channels. This situation is further investigated in [85] in relation to de Finetti's theorem [37], which is reformulated there in terms of multinomial channels forming a limit cone.

The multinomial channels do *not* commute with draw-add channels DA. For instance, for $\omega = \frac{1}{3}|a\rangle + \frac{2}{3}|b\rangle$ one has:

$$\begin{aligned}
mn[2](\omega) &= \tfrac{1}{9}\big|2|a\rangle\big\rangle + \tfrac{4}{9}\big|1|a\rangle + 1|b\rangle\big\rangle + \tfrac{4}{9}\big|2|b\rangle\big\rangle \\
mn[3](\omega) &= \tfrac{1}{27}\big|3|a\rangle\big\rangle + \tfrac{2}{9}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{4}{9}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{8}{27}\big|3|b\rangle\big\rangle.
\end{aligned}$$

But:

$$DA \gg\!= mn[2](\omega) = \tfrac{1}{9}\big|3|a\rangle\big\rangle + \tfrac{2}{9}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{2}{9}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{4}{9}\big|3|b\rangle\big\rangle.$$

There is one more point that we like to address. Since $mn[K](\omega)$ is a distribution, the sum over all draws $\sum_\varphi mn[K](\omega)(\varphi)$ equals one. But what if we restrict this sum to draws φ of certain colours only, that is, with $supp(\varphi) \subseteq S$, for a proper subset $S \subseteq supp(\omega)$? And what if we then let the size of these draws

K go to infinity? The result below describes what happens. It turns out that the same behaviour exists in the hypergeometric and Pólya cases, see Propositions 3.4.9.

Proposition 3.3.9. *Let $\omega \in \mathcal{D}(X)$ be given with a proper, non-empty subset $S \subseteq supp(\omega)$, so $S \neq \emptyset$ and $S \neq supp(\omega)$. For $K \in \mathbb{N}$, write:*

$$M_K \coloneqq \sum_{\varphi \in \mathcal{M}[K](S)} mn[K](\omega)(\varphi).$$

Then $M_K > M_{K+1}$ and $\lim_{K\to\infty} M_K = 0$.

Proof. Write $r \coloneqq \sum_{x\in S} \omega(x)$ in:

$$M_K = \sum_{\varphi \in \mathcal{M}[K](S)} mn[K](\omega)(\varphi) = \sum_{\varphi \in \mathcal{M}[K](S)} \binom{K}{\varphi} \cdot \prod_{x\in S} \omega(x)^{\varphi(x)} \overset{(1.27)}{=} r^K.$$

Since $0 < r < 1$ we get $M_K = r^K > r^{K+1} = M_{K+1}$ and $\lim_{K\to\infty} M_K = \lim_{K\to\infty} r^K = 0$. ❑

In this section we have assumed that an urn is given, namely as a distribution ω on colours, and we have looked at probabilities of multiple draws from the urn φ, as multisets. This will be turned around in the context of learning, see Chapter **??**. There we ask, among other things, the question: suppose that φ of size K is given, as 'data'; which distribution ω best fits φ, in the sense that it maximises the multinomial probability $mn[K](\omega)(\varphi)$? Perhaps unsurprisingly, the answer is: the distribution $Flrn(\varphi)$, obtained from φ via frequentist learning, see Proposition **??**.

Exercises

3.3.1 Let's throw a fair dice 12 times. What is the probability that each number appears twice? Show that it is $\frac{12!}{72^6}$.

3.3.2 This exercise illustrates naturality of the multinomial channel, see Corollary 3.3.2. Take sets $X = \{a, b, c\}$ and $Y = 2 = \{0, 1\}$ with function $f\colon X \to Y$ given by $f(a) = f(b) = 0$ and $f(c) = 1$. Consider the following distribution and multiset:

$$\omega = \tfrac{1}{4}|a\rangle + \tfrac{1}{2}|b\rangle + \tfrac{1}{4}|c\rangle \qquad \psi = 2|0\rangle + 1|1\rangle.$$

1 Show that:

$$mn[3]\Big(\mathcal{D}(f)(\omega)\Big)(\psi) = \tfrac{27}{64}.$$

2 Check that:

$$\mathcal{D}(\mathcal{N}[3](f))\Big(mn[3](\omega)\Big)(\psi) = mn[3](\omega)(2|a\rangle + 1|c\rangle) + mn[3](\omega)(1|a\rangle + 1|b\rangle + 1|c\rangle) + mn[3](\omega)(2|b\rangle + 1|c\rangle)$$

yields the same outcome $\frac{27}{64}$.

3.3.3 Check that:

$$mn[2]\Big(\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle\Big) = \tfrac{1}{4}\big|2|a\rangle\big\rangle + \tfrac{1}{2}\big|1|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|2|b\rangle\big\rangle.$$

Conclude that the multinomial map does not preserve uniform distributions.

3.3.4 Check that:

$$mn[1](\omega) = \sum_{x \in supp(\omega)} \omega(x)\,\big|1|x\rangle\big\rangle.$$

3.3.5 Use Exercise 1.6.3 to show that for $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](X)$ one has:

$$mn[K+L](\omega)(\varphi+\psi) = \frac{\binom{K+L}{K}}{\binom{\varphi+\psi}{\varphi}} \cdot mn[K](\omega)(\varphi) \cdot mn[L](\omega)(\psi),$$

and if φ, ψ have disjoint support, then:

$$mn[K+L](\omega)(\varphi+\psi) = \binom{K+L}{K} \cdot mn[K](\omega)(\varphi) \cdot mn[L](\omega)(\psi).$$

3.3.6 The aim of this exercise is to prove recurrence relations for multinomials: for each $\omega \in \mathcal{D}(X)$ and $\varphi \in \mathcal{N}[K](X)$ with $K > 0$ one has:

$$mn[K](\omega)(\varphi) = \sum_{x \in supp(\varphi)} \omega(x) \cdot mn[K-1](\omega)(\varphi - 1|x\rangle).$$

Here are two possible avenues:

1 use the recurrence relations (1.25) for multiset coefficients $\binom{\varphi}{}$;
2 show first:

$$\omega(x) \cdot mn[K-1](\omega)(\varphi - 1|x\rangle) = Flrn(\varphi)(x) \cdot mn[K](\omega)(\varphi).$$

Follow-up both avenues.

3.3.7 Prove the following items, in the style of Lemma 3.3.4, for a distribution $\omega \in \mathcal{D}(X)$ and a number $K > 1$.

1 For two elements $y \neq z$ in X,

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \cdot \varphi(z) = K \cdot (K-1) \cdot \omega(y) \cdot \omega(z).$$

2 For a single element $y \in X$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \cdot (\varphi(y) - 1) \;=\; K \cdot (K-1) \cdot \omega(y)^2.$$

3 Again, for a single element $y \in X$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y)^2 \;=\; K \cdot (K-1) \cdot \omega(y)^2 + K \cdot \omega(y).$$

Hint: Write $a^2 = a \cdot (a-1) + a$ and then use the previous item and Lemma 3.3.4.

3.3.8 1 Prove the following multivariate generalisation of Exercise 2.4.3, demonstrating that not only binomials, but more generally, multinomials are closed under convolution (2.17): for $K, L \in \mathbb{N}$,

$$\begin{array}{ccc} \mathcal{D}(X) \times \mathcal{D}(X) & \xrightarrow{mn[K] \otimes mn[L]} & \mathcal{N}[K](X) \times \mathcal{N}[L](X) \\ {\scriptstyle\Delta}\uparrow & & \downarrow{\scriptstyle +} \\ \mathcal{D}(X) & \xrightarrow{mn[K+L]} & \mathcal{N}[K+L](X) \end{array}$$

On the right, the sum + is the usual sum of multisets, promoted to a channel $\langle + \rangle$.

Hint: Use Exercises 1.6.5 and 2.5.7.

2 Conclude that K-sized draws can be reduced to parallel single draws, as in:

$$\begin{array}{ccc} \mathcal{D}(X)^K & \xrightarrow{mn[1]^K} & \mathcal{N}[1](X)^K \\ {\scriptstyle\Delta}\uparrow & & \downarrow{\scriptstyle +} \\ \mathcal{D}(X) & \xrightarrow{mn[K]} & \mathcal{N}[K](X) \end{array}$$

Notice that this is essentially Theorem 3.3.1 (2), via the isomorphism $\mathcal{M}[1](X) \cong X$.

3.3.9 Show that for a natural multiset φ of size K one has:

$$flat\Big(mn[K](Flrn(\varphi))\Big) \;=\; \varphi.$$

3.3.10 Check that multinomial channels do *not* commute with tensors, as in:

$$\begin{array}{ccc} \mathcal{D}(X) \times \mathcal{D}(Y) & \xrightarrow{mn[K] \otimes mn[L]} & \mathcal{M}[K](X) \times \mathcal{M}[L](Y) \\ {\scriptstyle\otimes}\downarrow & \neq & \downarrow{\scriptstyle\otimes} \\ \mathcal{D}(X \times Y) & \xrightarrow[mn[K \cdot L]]{} & \mathcal{M}[K \cdot L](X \times Y) \end{array}$$

3.3.11 Check that the following diagram does *not* commute.

$$\begin{array}{ccc} \mathcal{N}[K+1](X) \times \mathcal{N}[L+1](X) & \xrightarrow{DD \otimes DD} & \mathcal{N}[K](X) \times \mathcal{N}[L](X) \\ \Big\downarrow{\scriptstyle +} & \neq & \Big\downarrow{\scriptstyle +} \\ \mathcal{N}[K+L+2](X) & \xrightarrow[DD \circledcirc DD]{} & \mathcal{N}[K+L](X) \end{array}$$

3.3.12 Check that the multiset distribution $mltdst[K]_X$ from Exercise 2.1.7 equals:

$$mltdst[K]_X \;=\; mn[K](unif_X).$$

3.3.13 1 Check that accumulation does *not* commute with draw-and-delete, as in:

$$\begin{array}{ccc} X^{K+1} & \xrightarrow{\langle acc \rangle} & \mathcal{N}[K+1](X) \\ \Big\downarrow{\scriptstyle \pi} & \neq & \Big\downarrow{\scriptstyle DD} \\ X^K & \xrightarrow[\langle acc \rangle]{} & \mathcal{N}[K](X) \end{array}$$

2 Now define the probabilistic projection channel $ppr \colon X^{K+1} \nrightarrow X^K$ by:

$$\begin{array}{l} ppr(x_1, \ldots, x_{K+1}) \\ \quad := \displaystyle\sum_{1 \leq k \leq K+1} \frac{1}{K+1} \big| x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_{K+1} \big\rangle. \end{array} \tag{3.11}$$

Show that with ppr instead of π we do get a commuting diagram:

$$\begin{array}{ccc} X^{K+1} & \xrightarrow{\langle acc \rangle} & \mathcal{N}[K+1](X) \\ \Big\downarrow{\scriptstyle ppr} & & \Big\downarrow{\scriptstyle DD} \\ X^K & \xrightarrow{\langle acc \rangle} & \mathcal{N}[K](X) \end{array}$$

3 Conclude that we could have introduced DD via the definition principle of Proposition 3.1.1 as:

$$\begin{array}{ccccc} X^{K+1} & & \xrightarrow{\quad acc \quad} & & \mathcal{N}[K+1](X) \\ & \searrow^{ppr} & & & \Big\downarrow{\scriptstyle DD} \\ & & \mathcal{D}(X^K) & \xrightarrow[\mathcal{D}(acc)]{} & \mathcal{D}(\mathcal{N}[K](X)) \end{array}$$

4 Show that probabilistic projection also commutes with arrangement:

$$\begin{array}{ccc} X^{K+1} & \xleftarrow{arr} & \mathcal{N}[K+1](X) \\ \Big\downarrow{\scriptstyle ppr} & & \Big\downarrow{\scriptstyle DD} \\ X^K & \xleftarrow{arr} & \mathcal{N}[K](X) \end{array}$$

3.3.14 Show that the probabilistic projection channel *ppr* from the previous exercise makes the following diagram commute.

$$\begin{array}{ccc} \mathcal{D}(X)^{K+1} & \xrightarrow{\bigotimes} & X^{K+1} \\ {\scriptstyle ppr}\downarrow & & \downarrow{\scriptstyle ppr} \\ \mathcal{D}(X)^{K} & \xrightarrow{\bigotimes} & X^{K} \end{array}$$

3.3.15 1 Prove that $ppr \circ arr = \pi \circ arr$ in:

$$\mathcal{N}[K+1](X) \xrightarrow{arr} X^{K+1} \underset{\pi}{\overset{ppr}{\rightrightarrows}} X^{K}$$

2 Use this result to show that:

$$ppr \circ zip \circ (arr \otimes arr) \;=\; \pi \circ zip \circ (arr \otimes arr),$$

in a situation:

$$\begin{array}{ccc} \mathcal{N}[K+1](X) \times \mathcal{N}[K+1](Y) & \xrightarrow{arr \otimes arr} & X^{K+1} \times Y^{K+1} \\ & & {\scriptstyle zip}\downarrow \\ & & (X \times Y)^{K+1} \underset{\pi}{\overset{ppr}{\rightrightarrows}} (X \times Y)^{K} \end{array}$$

Hint: Use Proposition 3.2.2 (4).

3 Use the last item, (3.7), Exercises 2.2.6 and 3.3.13 (3), to show:

$$\begin{array}{ccc} \mathcal{N}[K+1](X) \times \mathcal{N}[K+1](Y) & \xrightarrow{DD \otimes DD} & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \\ {\scriptstyle mzip}\downarrow & & \downarrow{\scriptstyle mzip} \\ \mathcal{N}[K+1](X \times Y) & \xrightarrow{\quad DD \quad} & \mathcal{N}[K](X \times Y) \end{array}$$

3.3.16 Recall the set $\mathcal{D}_\infty(X)$ of discrete distributions from (2.8), with infinite (countable) support, see Exercise 2.1.11. For $\rho \in \mathcal{D}_\infty(\mathbb{N}_{>0})$ and $K \in \mathbb{N}$ define:

$$\underset{\infty}{mn}[K](\rho) \;\coloneqq\; \sum_{\varphi \in \mathcal{N}[K](\mathbb{N}_{>0})} \binom{K}{\varphi} \cdot \prod_{i \in supp(\varphi)} \rho(i)^{\varphi(i)} \,\big|\varphi\big\rangle.$$

1 Show that this yields a distribution in $\mathcal{D}_\infty(\mathcal{N}[K](\mathbb{N}_{>0}))$, *i.e.* that

$$\sum_{\varphi \in \mathcal{N}[K](\mathbb{N}_{>0})} \underset{\infty}{mn}[K](\rho)(\varphi) \;=\; 1.$$

2 Check that $DD \circ \underset{\infty}{mn}[K+1] = \underset{\infty}{mn}[K]$, like in Proposition 3.3.8.

3.4 The hypergeometric and Pólya channels

We have seen that a multinomial channel assigns probabilities to draws, with replacement. In addition, we have distinghuised two draws without replacement, namely the "-1" hypergeometric version where a drawn ball is actually removed from the urn, and the "+1" Pólya version where the drawn ball is returned to the urn together with an additional ball of the same colour. The additional ball has a strengthening effect that can capture situations with a cluster dynamics, like in the spread of contagious diseases [62] or the flow of tourists [108]. This section describes the main properties of these "-1" and "+1" draws. We have already briefly seen hypergeometric and Pólya distributions, namely in Example 2.1.1, in items (3) and (4).

Since drawn balls are removed or added in the hypergeometric and Pólya modes, the urn in question changes with each draw. It thus not a distribution, like in the multinomial case, but a multiset, say of size $L \in \mathbb{N}$. Draws are described as multisets of size K — with restriction $K \leq L$ in hypergeometric mode. The hypergeometric and Pólya channels thus takes the form:

$$\mathcal{N}[L](X) \underset{pl[K]}{\overset{hg[K]}{\rightrightarrows}} \mathcal{D}(\mathcal{N}[K](X)). \tag{3.12}$$

We first repeat from Example 2.1.1 (3) the definition of the hypergeometric channel. For a set X and a natural numbers $K \leq L$ we have, for multiset/urn $\psi \in \mathcal{N}[L](X)$,

$$hg[K](\psi) \coloneqq \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L}{K}} \big|\varphi\big\rangle = \sum_{\varphi \leq_K \psi} \frac{\prod_x \binom{\psi(x)}{\varphi(x)}}{\binom{L}{K}} \big|\varphi\big\rangle. \tag{3.13}$$

Recall that we write $\varphi \leq_K \psi$ for: $\|\varphi\| = K$ and $\varphi \leq \psi$, see Definition 1.5.1 (2). Lemma 1.6.2 shows that the probabilities in the above definition add up to one.

The Pólya channel resembles the above hypergeometric one, instead that multichoose $\left(\!\binom{}{}\!\right)$ is used instead of ordinary binomial coefficients $\binom{}{}$.

$$\begin{aligned} pl[K](\psi) &\coloneqq \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} \big|\varphi\big\rangle \\ &= \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \frac{\prod_{x \in supp(\psi)} \left(\!\binom{\psi(x)}{\varphi(x)}\!\right)}{\left(\!\binom{L}{K}\!\right)} \big|\varphi\big\rangle. \end{aligned} \tag{3.14}$$

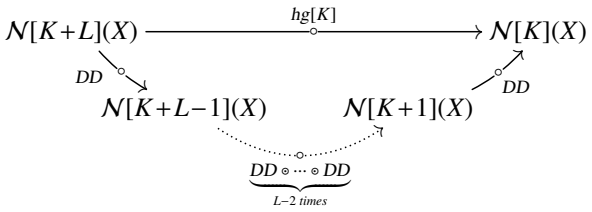
This yields a proper distribution by Proposition 1.7.3. In Section 3.5 we shall see how this distribution arises from iterated drawing and addition.

Pólya distributions have their own dynamics, different from hypergeometric distributions. One commonly refers to this well-studied approach as *Pólya urns*, see *e.g.* [115]. This section introduces the basics of such urns and the next section shows how they come about via iteration of a transition map, followed by marginalisation. Later on in Section **??**, within the continuous context, the Pólya channel shows up again, in relation to Dirichlet-multinomials and de Finetti's Theorem.

A subtle point in the above definitions (3.13) and (3.14) is that the draws φ should be multisets over the support $supp(\psi)$ of the urn ψ. Indeed, only balls that occur in the urn can be drawn. In the hypergeometric case this is handled via the requirement $\varphi \leq_K \psi$, which ensures an inclusion $supp(\varphi) \subseteq supp(\psi)$. In the Pólya case we ensure this inclusion by requiring that draws φ are in $\mathcal{N}[K](supp(\psi))$. Notice also that the product $\prod$ in (3.14) is restricted to range over $x \in supp(\psi)$ so that $\psi(x) > 0$. This is needed since $\binom{n}{m}$ is defined only for $n > 0$, see Definition 1.7.1 (2).

We start our analysis with hypergeometric channels. A basic fact is that they can be described in as iterations of draw-and-delete channels from Definition 2.2.4. As we shall see soon afterwards, this result has many useful consequences.

Theorem 3.4.1. *For $L, K \in \mathbb{N}$, the hypergeometric channel $\mathcal{N}[K+L](X) \multimap \mathcal{N}[K](X)$ equals L consecutive draw-and-delete's:*



To emphasise, this result says that the draws can be obtained via iterated draw-and-delete. The full picture emerges in Theorem 3.5.10 where we include the urn.

Proof. Write $\psi \in \mathcal{N}[K+L](X)$ for the urn. The proof proceeds by induction on the number of iterations L, starting with $L = 0$. Then $\varphi \leq_K \psi$ means $\varphi = \psi$. Hence:

$$hg[K](\psi) = \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{K+0}{K}} \big|\varphi\big\rangle = \frac{\binom{\psi}{\psi}}{\binom{K}{K}} \big|\psi\big\rangle = 1\big|\psi\big\rangle = DD^0(\psi) = DD^L(\psi).$$

For the induction step we use $\psi \in \mathcal{N}[K+(L+1)](X) = \mathcal{N}[(K+1)+L](X)$ and

$\varphi \leq_K \psi$. Then:

$$\begin{aligned}
DD^{L+1}(\psi)(\varphi) &= \Big(DD \gg\!= DD^L(\psi)\Big)(\varphi) \\
&= \sum_{\chi \in \mathcal{N}[K+1](X)} DD^L(\psi)(\chi) \cdot DD(\chi)(\varphi) \\
&= \sum_{y \in X} DD^L(\psi)(\varphi + 1|y\rangle) \cdot \frac{\varphi(y)+1}{K+1} \\
&\overset{\text{(IH)}}{=} \sum_{y,\, \varphi(y)<\psi(y)} \frac{\binom{\psi}{\varphi+1|y\rangle}}{\binom{K+L+1}{K+1}} \cdot \frac{\varphi(y)+1}{K+1} \\
&\overset{(*)}{=} \sum_{y,\, \varphi(y)<\psi(y)} \frac{(\psi(y)-\varphi(y)) \cdot \binom{\psi}{\varphi}}{(L+1) \cdot \binom{K+L+1}{K}} \quad \text{by Exercise 1.7.6 (1),(2)} \\
&= \frac{((K+L+1)-K) \cdot \binom{\psi}{\varphi}}{(L+1) \cdot \binom{K+L+1}{K}} \\
&= \frac{\binom{\psi}{\varphi}}{\binom{K+L+1}{K}} \\
&= hg[K](\psi)(\varphi). \qquad \blacksquare
\end{aligned}$$

From this result we can deduce many additional facts about hypergeometric distributions.

Corollary 3.4.2. *1 Hypergeometric channels* (3.12) *are natural in X.*

2 Frequentist learning form hypergeometric draws is like learning from the urn:

$$\begin{array}{ccc}
\mathcal{N}[N](X) & \xrightarrow{\quad hg[K] \quad} & \mathcal{N}[K](X) \\
& {\scriptstyle Flrn} \searrow \quad \swarrow {\scriptstyle Flrn} & \\
& X &
\end{array}$$

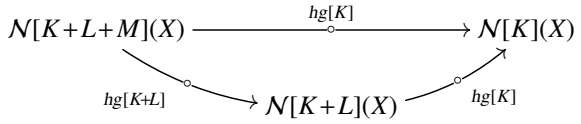
3 For $L \geq K$ one has:

$$\begin{array}{ccc}
\mathcal{N}[L](X) & \xleftarrow{\quad DD \quad} & \mathcal{N}[L+1](X) \\
& {\scriptstyle hg[K]} \searrow \quad \swarrow {\scriptstyle hg[K]} & \\
& \mathcal{N}[K](X) &
\end{array}$$

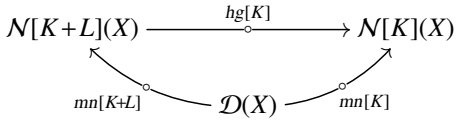
4 Also:

$$\begin{array}{ccc}
\mathcal{N}[K](X) & \xleftarrow{\quad DD \quad} & \mathcal{N}[K+1](X) \\
& {\scriptstyle hg[K]} \nwarrow \quad \nearrow {\scriptstyle hg[K+1]} & \\
& \mathcal{N}[L](X) &
\end{array}$$

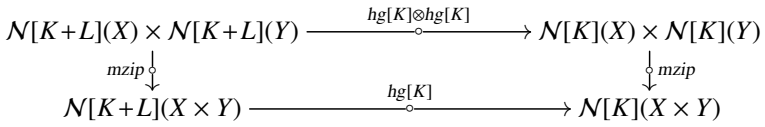
5 Hypergeometric channels compose, as in:



6 Hypergeometric and multinomial channels commute, as in:



7 Hypergeometric channels commute with multizip:



Proof. 1 By naturality of draw-and-delete, see Exercise 2.2.8.

2 By Theorem 2.3.3.

3 By Theorem 3.4.1.

4 Idem.

5 Similarly, since $DD^{L+M} = DD^L \circ DD^M$.

6 By Proposition 3.3.8.

7 By Exercise 3.3.15 (3). ❑

The distribution of small hypergeometric draws from a large urn looks like a multinomial distribution. This is intuitively clear, but will be made precise below.

Remark 3.4.3. When the urn from which we draw in hypergeometric mode is very large and the draw inolves only a small number of balls, the withdrawals do not really affect the urn. Hence in this case the hypergeometric distribution behaves like a multinomial distribution, where the urn (as distribution) is obtained via frequentist learning. This is elaborated below, where the urn ψ is

very large, in comparison to the draw $\varphi \leq_K \psi$.

$$\begin{aligned}
hg[K](\psi)(\varphi) \;=\; \frac{\binom{\psi}{\varphi}}{\binom{L}{K}} \;&=\; \frac{K!}{\prod_x \varphi(x)!} \cdot \prod\nolimits_x \frac{(L-K)!}{L!} \cdot \frac{\psi(x)!}{(\psi(x)-\varphi(x))!} \\
&=\; \multinom{\varphi} \cdot \prod\nolimits_x \frac{\psi(x) \cdot (\psi(x)-1) \cdot \ldots \cdot (\psi(x)-\varphi(x)+1)}{L \cdot (L-1) \cdot \ldots \cdot (L-K+1)} \\
&\overset{\text{big } \psi}{\approx}\; \multinom{\varphi} \cdot \prod\nolimits_x \frac{\psi(x)^{\varphi(x)}}{L^K} \\
&=\; \multinom{\varphi} \cdot \prod\nolimits_x \left(\frac{\psi(x)}{L}\right)^{\varphi(x)} \\
&=\; \multinom{\varphi} \cdot \prod\nolimits_x Flrn(\psi)(x)^{\varphi(x)} \\
&=\; mn[K](Flrn(\psi))(\varphi).
\end{aligned}$$

The essence of this observation is in Lemma 1.3.8.

We continue with Pólya distributions. Since hypergeometric draw channels arise from repeated draw-and-delete's one may expect that Pólya draw channels arise from repeted draw-and-add's. This is not the case. The connection will be clarified further in the next section. At this stage our first point below is that frequentist learning from its draws is the same as learning from the urn.

We shall first prove the following analogues of Lemma 3.3.4.

Lemma 3.4.4. *For an non-empty urn $\psi \in \mathcal{N}(X)$ and a fixed element $y \in X$,*

$$\sum_{\varphi \leq_K \psi} hg[K](\psi)(\varphi) \cdot \varphi(y) \;=\; K \cdot Flrn(\psi)(y), \tag{3.15}$$

And also:

$$\sum_{\varphi \in \mathcal{M}[K](supp(\psi))} pl[K](\psi)(\varphi) \cdot \varphi(y) \;=\; K \cdot Flrn(\psi)(y). \tag{3.16}$$

Proof. We use Exercises 1.7.5 and 1.7.6 in the marked equation $\overset{(*)}{=}$ below. We start with equation (3.15), for which we assume $K \leq L$.

$$\begin{aligned}
\sum_{\varphi \leq_K \psi} hg[K](\psi)(\varphi) \cdot \varphi(y) \;&=\; \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi} \cdot \varphi(y)}{\binom{L}{K}} \\
&\overset{(*)}{=}\; \sum_{\varphi \leq_K \psi} \frac{\psi(y) \cdot \binom{\psi - 1|y\rangle}{\varphi - 1|y\rangle}}{\frac{L}{K} \cdot \binom{L-1}{K-1}} \\
&=\; K \cdot \frac{\psi(y)}{L} \cdot \sum_{\varphi \leq_{K-1} \psi - 1|y\rangle} \frac{\binom{\psi-1|y\rangle}{\varphi}}{\binom{L-1}{K-1}} \\
&=\; K \cdot Flrn(\psi)(y).
\end{aligned}$$

Similarly we obtain (3.16) via the additional use of Exercise 1.7.5.

$$\begin{aligned}
\sum_{\varphi\in\mathcal{M}[K](supp(\psi))} pl[K](\psi)(\varphi)\cdot\varphi(y) &= \sum_{\varphi\in\mathcal{M}[K](supp(\psi))} \frac{\binom{\psi}{\varphi}\cdot\varphi(y)}{\binom{L}{K}} \\
&\overset{(*)}{=} \sum_{\varphi\in\mathcal{M}[K](supp(\psi))} \frac{\psi(y)\cdot\binom{\psi+1|y\rangle}{\varphi-1|y\rangle}}{\frac{L}{K}\cdot\binom{L+1}{K-1}} \\
&= K\cdot\frac{\psi(y)}{L}\cdot \sum_{\varphi\in\mathcal{M}[K-1](supp(\psi+1|y\rangle))} \frac{\binom{\psi-1|y\rangle}{\varphi}}{\binom{L+1}{K-1}} \\
&= K\cdot Flrn(\psi)(y). \qquad \blacksquare
\end{aligned}$$

Proposition 3.4.5. *Let $K > 0$.*

1 Frequentist learning and Pólya satisfy the following equation:

$$\begin{array}{ccc}
\mathcal{N}_*(X) & \xrightarrow{\ pl[K]\ } & \mathcal{N}[K](X) \\
& {\scriptstyle Flrn}\searrow \quad \swarrow{\scriptstyle Flrn} & \\
& X &
\end{array}$$

2 Doing a draw-and-add before Pólya has no effect: for $L > 0$,

$$\begin{array}{ccc}
\mathcal{N}[L](X) & \xrightarrow{\ DA\ } & \mathcal{N}[L+1](X) \\
& {\scriptstyle pl[K]}\searrow \quad \swarrow{\scriptstyle pl[K]} & \\
& \mathcal{N}[K](X) &
\end{array}$$

This second point is the Pólya analogue of Theorem 3.4.2 (3), with draw-and-add instead of draw-and-delete.

Proof. 1 For a multiset/urn $\psi \in \mathcal{N}(X)$ with $\|\psi\| = L > 0$, and for $x \in X$,

$$\begin{aligned}
\Big(Flrn \gg\!= pl[K](\psi)\Big)(x) &= \sum_{\varphi\in\mathcal{N}[K](supp(\psi))} Flrn(\varphi)(x)\cdot pl[K](\psi)(\varphi) \\
&= \frac{1}{K}\cdot \sum_{\varphi\in\mathcal{N}[K](supp(\psi))} \varphi(x)\cdot pl[K](\psi)(\varphi) \\
&\overset{(3.16)}{=} Flrn(\psi)(x).
\end{aligned}$$

2 We use Exercises 1.7.5 and 1.7.6 in the marked equation $\overset{(*)}{=}$ below. For urn

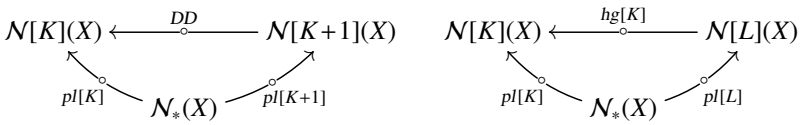
$\psi \in \mathcal{N}[L](X)$ and draw $\varphi \in \mathcal{N}[K](X)$,

$$\begin{aligned}
\Big(pl[K] \circ DA\Big)(\psi)(\varphi) &= \sum_{x \in supp(\psi)} \frac{\psi(x)}{L} \cdot pl[K](\psi + 1|x\rangle)(\varphi) \\
&= \sum_{x \in supp(\psi)} \frac{\psi(x)}{L} \cdot \frac{\left(\!\binom{\psi+1|x\rangle}{\varphi}\!\right)}{\left(\!\binom{L+1}{K}\!\right)} \\
&\overset{(*)}{=} \sum_{x \in supp(\psi)} \frac{\psi(x) + \varphi(x)}{L + K} \cdot \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} \\
&= \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} \\
&= pl[K](\psi)(\varphi). \qquad \text{❑}
\end{aligned}$$

Next we look at interaction of the Pólya channel with draw-and-delete — like in Theorem 3.4.2 (4) and then also with hypergeometric channels.

The next point illustrates a relationship between Pólya and the draw-and-add channel DA from Definition 2.2.4.

Theorem 3.4.6. *For each K and $L \geq K$ the triangles below commute.*



Proof. We concentrate on commutation of the triangle on the left since it implies commutation of the on on the right by Theorem 3.4.1. The marked equation $\overset{(*)}{=}$ below indicates the use of Exercises 1.7.5 and 1.7.6. For urn $\psi \in \mathcal{N}[L](X)$ of size $L > 0$, and for $\varphi \in \mathcal{N}[K](X)$,

$$\begin{aligned}
\Big(DD \circ pl[K+1]\Big)(\psi)(\varphi) &= \sum_{\chi \in \mathcal{N}[K+1](supp(\psi))} DD(\chi)(\varphi) \cdot pl[K+1](\psi)(\chi) \\
&= \sum_{x \in supp(\psi)} \frac{\varphi(x)+1}{K+1} \cdot pl[K+1](\psi)(\varphi + 1|x\rangle) \\
&= \sum_{x \in supp(\psi)} \frac{\varphi(x)+1}{K+1} \cdot \frac{\left(\!\binom{\psi}{\varphi+1|x\rangle}\!\right)}{\left(\!\binom{L}{K+1}\!\right)} \\
&\overset{(*)}{=} \sum_{x \in supp(\psi)} \frac{\psi(x) + \varphi(x)}{L + K} \cdot \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} \\
&= pl[K](\psi). \qquad \text{❑}
\end{aligned}$$

Later on we shall see that the Pólya channel factors through the multinomial channel, see Exercise 6.4.5 and **??**. The Pólya channel does not commute with multizip.

Earlier in Proposition 3.3.7 we have seen an 'average' result for multinomials. There are similar results for hypergeometric and Pólya distributions.

Proposition 3.4.7. *Let $\psi \in \mathcal{N}(X)$ be an urn of size $\|\psi\| = L \geq 1$.*

1 For $L \geq K \geq 1$,

$$\mathit{flat}(\mathit{hg}[K](\psi)) = \sum_{\varphi \leq_K \psi} \mathit{hg}[K](\psi)(\varphi) \cdot \varphi = \frac{K}{L} \cdot \psi = K \cdot \mathit{Flrn}(\psi).$$

2 Similarly, for $K \geq 1$,

$$\mathit{flat}(\mathit{pl}[K](\psi)) = \sum_{\varphi \in \mathcal{N}[K](\mathit{supp}(\psi))} \mathit{pl}[K](\psi)(\varphi) \cdot \varphi = \frac{K}{L} \cdot \psi = K \cdot \mathit{Flrn}(\psi).$$

Proof. In both cases we rely on Lemma 3.4.4. For the first item we get:

$$\begin{aligned} \sum_{\varphi \leq_K \psi} \mathit{hg}[K](\psi)(\varphi) \cdot \varphi &= \sum_{x \in X} \left(\sum_{\varphi \leq_K \psi} \mathit{hg}[K](\psi)(\varphi) \cdot \varphi(x) \right) \big| x \big\rangle \\ &\overset{(3.15)}{=} \sum_{x \in X} K \cdot \mathit{Flrn}(\psi)(x) \, \big| x \big\rangle \\ &= K \cdot \mathit{Flrn}(\psi). \end{aligned}$$

In the second, Pólya case we proceed analogously:

$$\begin{aligned} \sum_{\varphi \in \mathcal{N}[K](\mathit{supp}(\psi))} \mathit{pl}[K](\psi)(\varphi) \cdot \varphi &= \sum_{x \in X} \left(\sum_{\varphi \in \mathcal{N}[K](\mathit{supp}(\psi))} \mathit{pl}[K](\psi)(\varphi) \cdot \varphi(x) \right) \big| x \big\rangle \\ &\overset{(3.16)}{=} \sum_{x \in X} K \cdot \mathit{Flrn}(\psi)(x) \, \big| x \big\rangle \\ &= K \cdot \mathit{Flrn}(\psi). \end{aligned}$$

❑

There is another point of analogy with multinomial distributions that we wish to elaborate, namely the (limit) behaviour when balls of specific colours only are drawn, see Proposition 3.3.9. In the hypergeometric case it does not make sense to look at limit behaviour since after a certain number of steps the urn is empty. But in the Pólya case one can continue drawing indefinitely, making not trivial what happens in the limit. We use the following auxiliary result[1].

[1] With thanks to Bas and Bram Westerbaan for help.

Lemma 3.4.8. *Let $0 < N < M$ be given. Define for $n \in \mathbb{N}$,*

$$a_n := \prod_{i<n} \frac{N+i}{M+i}.$$

Then: $\lim_{n\to\infty} a_n = 0$.

Proof. We switch to the (natural) logarithm ln and prove the equivalent statement $\lim_{n\to\infty} \ln(a_n) = -\infty$. We use that the logarithm turns products into sums, see Exercise 1.2.2, and that the derivative of $\ln(x)$ is $\frac{1}{x}$. Then:

$$\begin{aligned}
\ln(a_n) \;=\; \sum_{i<n} \ln\Big(\frac{N+i}{M+i}\Big) \;&=\; \sum_{i<n} \ln(N+i) - \ln(M+i) \\
&=\; -\sum_{i<n} \int_{N+1}^{M+1} \frac{1}{x}\,\mathrm{d}x \\
&\overset{(*)}{\leq} -\sum_{i<n} \frac{(M+i)-(N+i)}{M+i} \\
&=\; (N-M)\cdot \sum_{i<n} \frac{1}{M+i} \\
&=\; (N-M)\cdot \sum_{i<n} \frac{1}{M+i}.
\end{aligned}$$

It is well-known that the harmonic series $\sum_{n>0} \frac{1}{n}$ is infinite. Since $M > N$ the above sequence $\ln(a_n)$ thus goes to $-\infty$.

The validity of the marked inequality $\leq$ follows from an inspection of the graph of the function $\frac{1}{x}$: the integral from $N+i$ to $M+i$ is the surface under $\frac{1}{x}$ between the points $N+i < M+i$. Since $\frac{1}{x}$ is a decreasing function, this surface is bigger than the rectangle with height $\frac{1}{M+i}$ and length $(M+i)-(N+i)$. ❑

Proposition 3.4.9. *Consider an urn $\psi \in \mathcal{N}[L](X)$ with a proper non-empty subset $S \subseteq supp(\psi)$.*

1 Write for $K \leq L$,

$$H_K := \sum_{\varphi\in\mathcal{N}[K](S),\, \varphi\leq\psi} hg[K](\psi)(\varphi).$$

Then $H_K > H_{K+1}$; this stops at $K = L$, when the urn is empty.

2 Now write, for arbitrary $K \in \mathbb{N}$,

$$P_K := \sum_{\varphi\in\mathcal{N}[K](S)} pl[K](\psi)(\varphi).$$

Then $P_K > P_{K+1}$ and $\lim_{K\to\infty} P_K = 0$.

Proof. We write $L_S \coloneqq \sum_{x\in S} \psi(x)$ is for the number of balls in the urn whose colour is in S.

1 By separating S and its complement $\neg S$ we can write, via Vandermonde's formula from Lemma 1.6.2,

$$H_K = \sum_{\varphi\in\mathcal{N}[K](S),\, \varphi\leq\psi} \frac{\left(\prod_{x\in S}\binom{\psi(x)}{\varphi(x)}\right)\cdot\left(\prod_{x\notin S}\binom{\psi(x)}{0}\right)}{\binom{L}{K}} = \frac{\binom{L_S}{K}}{\binom{L}{K}} = \frac{L_S\,!}{L!}\cdot\frac{(L-K)!}{(L_S-K)!}.$$

Using a similar description for H_{K+1} we get:

$$\begin{aligned} H_K > H_{K+1} &\iff \frac{L_S\,!}{L!}\cdot\frac{(L-K)!}{(L_S-K)!} > \frac{L_S\,!}{L!}\cdot\frac{(L-(K+1))!}{(L_S-(K+1))!} \\ &\iff \frac{L-K}{L_S-K} > 1. \end{aligned}$$

The latter holds because $L > L_S$, since S is a proper subset of $supp(\psi)$.

2 In the Pólya case we get, as before, but now via the Vandermonde formula for multichoose, see Proposition 1.7.3,

$$P_K = \frac{\left(\!\binom{L_S}{K}\!\right)}{\left(\!\binom{L}{K}\!\right)} = \frac{(L-1)!}{(L_S-1)!}\cdot\frac{(L_S+K-1)!}{(L+K-1)!}.$$

We define:

$$\begin{aligned} a_K \coloneqq \frac{P_{K+1}}{P_K} &= \frac{(L-1)!}{(L_S-1)!}\cdot\frac{(L_S+(K+1)-1)!}{(L+(K+1)-1)!}\cdot\frac{(L_S-1)!}{(L-1)!}\cdot\frac{(L+K-1)!}{(L_S+K-1)!} \\ &= \frac{L_S+K}{L+K} < 1, \qquad \text{since } L_S < L. \end{aligned}$$

Thus $P_{K+1} = a_K \cdot P_K < P_K$ and also:

$$P_K = a_{K-1}\cdot P_{K-1} = a_{K-1}\cdot a_{K-2}\cdot P_{K-2} = \cdots = a_{K-1}\cdot a_{K-2}\cdot\ldots\cdot a_0\cdot P_0.$$

Our goal is to prove $\lim_{K\to\infty} P_K = 0$. This follows from $\lim_{K\to\infty}\prod_{i<K} a_i = 0$, which we obtain from Lemma 3.4.8. ❑

Exercises

3.4.1 Consider an urn with 5 red, 6 green, and 8 blue balls. Suppose 6 balls are drawn from the urn, resulting in two balls of each colour.

1 Describe both the urn and the draw as a multiset.

Show that the probability of the six-ball draw is:

2 $\frac{5184000}{47045881} \approx 0.11$, when balls are drawn one by one and are replaced before each next single-ball draw;

3 $\frac{50}{323} \approx 0.15$, when the drawn balls are deleted;
4 $\frac{405}{4807} \approx 0.08$ when the drawn balls are replaced and each time an extra ball of the same colour is added.

3.4.2 Draw-delete preserves hypergeometric and Pólya distributions, see Corollary 3.4.2 (4) and Theorem 3.4.6. Check that draw-add does not preserve hypergeometric or Pólya distributions, for instance by checking that for $\varphi = 3|a\rangle + 1|b\rangle$,

$$\begin{aligned} hg[2](\varphi) &= \tfrac{1}{2}\big|2|a\rangle\big\rangle + \tfrac{1}{2}\big|1|a\rangle + 1|b\rangle\big\rangle \\ hg[3](\varphi) &= \tfrac{1}{4}\big|3|a\rangle\big\rangle + \tfrac{3}{4}\big|2|a\rangle + 1|b\rangle\big\rangle \\ DA \gg\!= hg[2](\varphi) &= \tfrac{1}{2}\big|3|a\rangle\big\rangle + \tfrac{1}{4}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|1|a\rangle + 2|b\rangle\big\rangle, \end{aligned}$$

And:

$$\begin{aligned} &pl[2](\varphi) \\ &= \tfrac{3}{5}\big|2|a\rangle\big\rangle + \tfrac{3}{10}\big|1|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{10}\big|2|b\rangle\big\rangle \\ &pl[3](\varphi) \\ &= \tfrac{1}{2}\big|3|a\rangle\big\rangle + \tfrac{3}{10}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{3}{20}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{20}\big|3|b\rangle\big\rangle \\ &DA \gg\!= pl[2](\varphi) \\ &= \tfrac{3}{5}\big|3|a\rangle\big\rangle + \tfrac{3}{20}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{3}{20}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{10}\big|3|b\rangle\big\rangle \end{aligned}$$

3.4.3 Let X be a finite set, say of size n. Write $\mathbf{1} = \sum_{x \in X} 1|x\rangle$ for the multiset of single occurrences of elements of X. Show that for each $K \geq 0$,

$$pl[K](\mathbf{1}) = unif_{\mathcal{N}[K](X)}.$$

3.4.4 1 Give a direct proof of Theorem 3.4.1 for $L = 1$.
2 Elaborate also the case $K = 1$ in Theorem 3.4.1 and show that in that case $DD^L(\psi) = Flrn(\psi)$, at least when we identify the (isomorphic) sets $\mathcal{N}[1](X)$ and X.

3.4.5 Let $unif_L \in \mathcal{D}(\mathcal{N}[L](X))$ be the uniform distribution on multisets of size L, from Exercise 2.2.9, where X is a finite set. Show that $hg[K] \gg\!= unif_L = unif_K$, for $K \leq L$.

3.4.6 Recall from Remark 3.4.3 that small hypergeometric draws from a large urn can be described multinomially. Extend this in the following way. Let $\sigma \in \mathcal{D}(\mathcal{N}[L](X))$ be a distribution of large urns. Argue that for small numbers K,

$$hg[K] \gg\!= \sigma \approx mn[K] \gg\!= \mathcal{D}(Flrn)(\sigma).$$

3.4.7 Prove in analogy with Exercise 3.3.7, for an urn $\psi \in \mathcal{N}[L](X)$ and elements $y, z \in X$, the following points.

1 When $y \neq z$ and $K \leq L$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} hg[K](\psi)(\varphi) \cdot \varphi(y) \cdot \varphi(z) = K \cdot (K-1) \cdot Flrn(\psi)(y) \cdot \frac{\psi(z)}{L-1}.$$

2 When $K \leq L$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} hg[K](\psi)(\varphi) \cdot \varphi(y)^2 = K \cdot Flrn(\psi)(y) \cdot \frac{(K-1) \cdot \psi(y) + (L-K)}{L-1}.$$

3 And in the Pólya case, when $y \neq z$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} pl[K](\psi)(\varphi) \cdot \varphi(y) \cdot \varphi(z) = K \cdot (K-1) \cdot Flrn(\psi)(y) \cdot \frac{\psi(z)}{L+1}.$$

4 Finally,

$$\sum_{\varphi \in \mathcal{N}[K](X)} pl[K](\psi)(\varphi) \cdot \varphi(y)^2 = K \cdot Flrn(\psi)(y) \cdot \frac{(K-1) \cdot \psi(y) + (L+K)}{L+1}.$$

3.4.8 Use Theorem 3.4.1 to prove the following two recurrence relations for hypgeometric distributions.

$$hg[K](\psi) = \sum_{x \in supp(\psi)} Flrn(\psi)(x) \cdot hg[K-1](\psi - 1|x\rangle)$$

$$hg[K](\psi)(\varphi) = \sum_{x} \frac{\varphi(x) + 1}{K+1} \cdot hg[K-1](\psi)(\varphi + 1|x\rangle)$$

3.4.9 Fix numbers $N, M \in \mathbb{N}$ and write $\psi = N|0\rangle + M|1\rangle$ for an urn with N balls of colour 0 and M of colour 1. Let $n \leq N$. Show that:

$$\sum_{0 \leq m \leq M} hg[n+m](\psi)(n|0\rangle + m|1\rangle) = \frac{N+M+1}{N+1}.$$

Hint: Use Exercise 1.7.3. Notice that the right-hand side does not depend on n.

3.4.10 This exercise elaborates that draws from an urn excluding one particular colour can be expressed in binary form. This works for all three modes of drawing: multinomial, hypergeometric, and Pólya.

Let X be a set with at least two elements, and let $x \in X$ be an arbitrary but fixed element. We write $x^\perp$ for an element not in X. Assume $k \leq K$.

1 For $\omega \in \mathcal{D}(X)$ with $x \in \mathit{supp}(\omega)$, use the Multinomial Theorem (1.27) to show that:

$$\begin{aligned}&\sum_{\varphi \in \mathcal{N}[K-k](X-x)} mn[K](\omega)(k|x\rangle + \varphi) \\ &= mn[K]\Big(\omega(x)|x\rangle + (1-\omega(x))|x^{\perp}\rangle\Big)\Big(k|x\rangle + (K-k)|x^{\perp}\rangle\Big) \\ &= bn[K](\omega(x))(k).\end{aligned}$$

2 Prove, via Vandermonde's formula, that for be an urn ψ of size $L \geq K$ one has:

$$\begin{aligned}&\sum_{\varphi \leq_{K-k} \psi - \psi(x)|x\rangle} hg[K](\psi)(k|x\rangle + \varphi) \\ &= hg[K]\Big(\psi(x)|x\rangle + (L-\psi(x))|x^{\perp}\rangle\Big)\Big(k|x\rangle + (K-k)|x^{\perp}\rangle\Big).\end{aligned}$$

3 Show again for an urn ψ,

$$\begin{aligned}&\sum_{\varphi \in \mathcal{N}[K-k](\mathit{supp}(\psi)-x)} pl[K](\psi)(k|x\rangle + \varphi) \\ &= pl[K]\Big(\psi(x)|x\rangle + (L-\psi(x))|x^{\perp}\rangle\Big)\Big(k|x\rangle + (K-k)|x^{\perp}\rangle\Big).\end{aligned}$$

3.5 Iterated drawing from an urn

This section deals with the iterative aspects of drawing from an urn. It makes the informal descriptions in the introduction to this chapter precise, for six different forms of drawing: in "-1" or "0" or "+1" mode, and with ordered or unordered draws. The iteration involved will be formalised via a special monad, combining the multiset and the writer monad. This infrastructure for iteration will be set up first. Subsequently, ordered and unordered draws will be described in separate subsections.

We recall the transition operation (3.2) from the beginning of this chapter:

$$\mathit{Urn} \longmapsto \Big(\mathit{multiple\text{-}colours},\ \mathit{Urn}'\Big)$$

mapping an urn to a distribution over pairs of multiple draws and remaining urn. The actual mapping from urns to draws would then arise as first marginal of this transition channel. Concretely, the above transition map, used for de-

scribing our intuition, takes one of the following six forms.

mode	ordered	unordered
0	$\mathcal{D}(X) \xrightarrow{Ot_0} \mathcal{L}(X) \times \mathcal{D}(X)$	$\mathcal{D}(X) \xrightarrow{Ut_0} \mathcal{N}(X) \times \mathcal{D}(X)$
-1	$\mathcal{N}_*(X) \xrightarrow{Ot_-} \mathcal{L}(X) \times \mathcal{N}_*(X)$	$\mathcal{N}_*(X) \xrightarrow{Ut_-} \mathcal{N}(X) \times \mathcal{N}_*(X)$
+1	$\mathcal{N}(X) \xrightarrow{Ot_+} \mathcal{L}(X) \times \mathcal{N}(X)$	$\mathcal{N}_*(X) \xrightarrow{Ut_+} \mathcal{N}(X) \times \mathcal{N}_*(X)$

(3.17)

Notice that the draw maps in Table 3.3 in the introduction are written as channels $O_0 \colon \mathcal{D}(X) \nrightarrow \mathcal{N}[K](X)$, where the above table contains the corresponding *transition* maps, with an extra 't' in the name, as in $Ot_0 \colon \mathcal{D}(X) \nrightarrow \mathcal{N}(X) \times \mathcal{D}(X)$.

In each case, the drawn elements are accumulated in the left product-component of the codomain of the transition map. They are organised as lists, in $\mathcal{L}(X)$, in the ordered case, and as multisets, in $\mathcal{N}(X)$ in the unordered case. As we have seen before, in the scenarios with deletion/addition the urn is a multiset, but with replacement it is a distribution.

The crucial observation is that the list and multiset data types that we use for accumulating drawn elements are monoids, as we have observed early on, in Lemmas 1.2.2 and 1.4.2. In general, for a monoid M, the mapping $X \mapsto M \times X$ is a monad, called the writer monad, see Lemma 1.9.1. It turns out that the combination of the writer monad with the distribution monad $\mathcal{D}$ is again a monad. This forms the basis for iterating transitions, via Kleisli composition of this combined monad. We relegate the details of the monad's flatten operation to Exercise 3.5.6 below, and concentrate on the unit and Kleisli composition involved.

Lemma 3.5.1. *Let $M = (M, 0, +)$ be a monoid. The mapping $X \mapsto \mathcal{D}(M \times X)$ is a monad on* **Sets***, with* $\mathit{unit} \colon X \to \mathcal{D}(M \times X)$ *given by:*

$$\mathit{unit}(x) \;=\; 1|0, x\rangle \qquad \textit{where } 0 \in M \textit{ is the zero element.}$$

For Kleisli maps $f \colon A \to \mathcal{D}(M \times B)$ and $g \colon B \to \mathcal{D}(M \times C)$ there is the Kleisli composition $g \mathbin{\odot} f \colon A \to \mathcal{D}(M \times C)$ given by:

$$(g \mathbin{\odot} f)(a) \;=\; \sum_{m,m',c} \Big(\textstyle\sum_b f(a)(m,b) \cdot g(b)(m',c) \Big) \big| m + m', c \big\rangle. \tag{3.18}$$

Notice the occurence of the sum + of the monoid M in the first component of the ket $\big|-,-\big\rangle$ in (3.18). When M is the list monoid, this sum is the (non-commutative) concatenation ++ of lists, producing an ordered list of drawn elements. When M is the multiset monoid, this sum is the (commutative) +

$$\begin{array}{ll}
\mathcal{D}(X) \xrightarrow{Ot_0} \mathcal{D}(\mathcal{L}(X) \times \mathcal{D}(X)) & \mathcal{N}(X) \xrightarrow{Ut_0} \mathcal{D}(\mathcal{N}(X) \times \mathcal{N}(X)) \\
\omega \longmapsto \sum_{x \in supp(\omega)} \omega(x) \big| [x], \omega \big\rangle & \omega \longmapsto \sum_{x \in supp(\omega)} \omega(x) \big| 1|x\rangle, \omega \big\rangle \\
\mathcal{N}_*(X) \xrightarrow{Ot_-} \mathcal{D}(\mathcal{L}(X) \times \mathcal{N}(X)) & \mathcal{N}_*(X) \xrightarrow{Ut_-} \mathcal{D}(\mathcal{N}(X) \times \mathcal{N}(X)) \\
\psi \longmapsto \sum_{x \in supp(\psi)} \frac{\psi(x)}{\|\psi\|} \big| [x], \psi - 1|x\rangle \big\rangle & \psi \longmapsto \sum_{x \in supp(\psi)} \frac{\psi(x)}{\|\psi\|} \big| 1|x\rangle, \psi - 1|x\rangle \big\rangle \\
\mathcal{N}_*(X) \xrightarrow{Ot_+} \mathcal{D}(\mathcal{L}(X) \times \mathcal{N}_*(X)) & \mathcal{N}_*(X) \xrightarrow{Ut_+} \mathcal{D}(\mathcal{N}(X) \times \mathcal{N}_*(X)) \\
\psi \longmapsto \sum_{x \in supp(\psi)} \frac{\psi(x)}{\|\psi\|} \big| [x], \psi + 1|x\rangle \big\rangle & \psi \longmapsto \sum_{x \in supp(\psi)} \frac{\psi(x)}{\|\psi\|} \big| 1|x\rangle, \psi + 1|x\rangle \big\rangle
\end{array}$$

Figure 3.1 Definitions of the six transition channels in Table (3.17) for drawing a single element from an urn. In the "ordered" column on the left the list monoid $\mathcal{L}(X)$ is used, where in the "unordered" columnn the (commutative) multiset monoid $\mathcal{N}(X)$ occurs. In the first row the urn (as distribution ω) remains unchanged, whereas in the second and (resp. third) row the drawn elements x is removed (resp. added) to the urn ψ. Implicitly it is assumed that the multiset ψ is non-empty.

of multisets, so that the accumulation of drawn elements yields a multiset, in which the order of elements is irrelevant.

If we have an 'endo' Kleisli map for the combined monad of Lemma 3.5.1, of the form $t\colon A \to \mathcal{D}(M \times A)$, we can iterate it K times, giving $t^K\colon A \to \mathcal{D}(M \times A)$. This iteration is defined via the above unit and Kleisli composition:

$$t^0 = unit \qquad \text{and} \qquad t^{K+1} = t^K \circ t = t \circ t^K.$$

Now that we know how to iterate, we need the actual transition maps that can be iterated, that is, we need concrete definitions of the transition channels in Table (3.17). They are given in Figure 3.1. In the subsections below we analyse what iteration means for these six channels. Subsequently, we can describe the associated K-sized draw channels, as in Table 3.3, as first projection $\pi_1 \circ t^K$, going from urns to drawn elements.

3.5.1 Ordered draws from an urn

We start to look at the upper two 'ordered' transition channels in the column on the left in Figure 3.1. Towards a general formula for their iteration, let's look first at the easiest case, namely ordered transition $Ot_0\colon \mathcal{D}(X) \to \mathcal{D}(\mathcal{L}(X) \times$

$\mathcal{D}(X)$) with replacement. By definition we have as first iteration.

$$Ot_0^1(\omega) \;=\; Ot_0(\omega) \;=\; \sum_{x_1 \in supp(\omega)} \omega(x_1) \,\big|\, [x_1], \omega \big\rangle.$$

Accumulation of drawn elements in the first coordinate of $\big|-,-\big\rangle$ starts in the second iteration:

$$\begin{aligned} Ot_0^2(\omega) \;&=\; Ot_0 \gg\!= Ot_0(\omega) \\ &=\; \sum_{\ell \in \mathcal{L}(X),\, x_1 \in supp(\omega)} \omega(x_1) \cdot Ot_0(\omega)(\ell, \omega) \,\big|\, [x_1] +\!\!+ \ell, \omega \big\rangle \\ &=\; \sum_{x_1, x_2 \in supp(\omega)} \omega(x_1) \cdot \omega(x_2) \,\big|\, [x_1] +\!\!+ [x_2], \omega \big\rangle \\ &=\; \sum_{x_1, x_2 \in supp(\omega)} (\omega \otimes \omega)(x_1, x_2) \,\big|\, [x_1, x_2], \omega \big\rangle. \end{aligned}$$

The formula for subsequent iterations is beginning to appear.

Theorem 3.5.2. *Consider in Figure 3.1 the ordered-transition-with-replacement channel $Ot_0 \colon \mathcal{D}(X) \nrightarrow \mathcal{L}(X) \times \mathcal{D}(X)$, with distribution $\omega \in \mathcal{D}(X)$.*

1 Iterating $K \in \mathbb{N}$ times yields:

$$Ot_0^K(\omega) \;=\; \sum_{\vec{x} \in X^K} \omega^K(\vec{x}) \,\big|\, \vec{x}, \omega \big\rangle.$$

2 The associated K-draw channel $O_0[K] \coloneqq \pi_1 \mathrel{\circ} Ot_0^K \colon \mathcal{D}(X) \nrightarrow X^K$ satisfies

$$O_0[K](\omega) \;=\; \omega^K \;=\; iid[K](\omega),$$

where iid is the identical and independent channel from (2.20). ❑

The situation for ordered transition with deletion is less straightforward. We look at two iterations explicitly, starting from a multiset $\psi \in \mathcal{N}(X)$.

$$\begin{aligned} Ot_-^1(\psi) \;&=\; \sum_{x_1 \in supp(\psi)} \frac{\psi(x_1)}{\|\psi\|} \,\big|\, x_1, \psi - 1|\, x_1 \,\big\rangle \big\rangle \\ Ot_-^2(\psi) \;&=\; Ot_- \gg\!= Ot_-(\psi) \\ &=\; \sum_{\substack{x_1 \in supp(\psi), \\ x_2 \in supp(\psi - 1|\, x_1 \,\rangle}} \frac{\psi(x_1)}{\|\psi\|} \cdot \frac{(\psi - 1|\, x_1 \,\rangle)(x_2)}{\|\psi\| - 1} \,\big|\, x_1, x_2, \psi - 1|\, x_1 \,\rangle - 1|\, x_2 \,\big\rangle \big\rangle. \end{aligned}$$

Etcetera. We first collect some basic observations in an auxiliary result.

Lemma 3.5.3. *Let $\psi \in \mathcal{N}[L](X)$ be a multiset / urn of size $L = \|\psi\|$.*

1 Iterating $K \leq L$ times satisfies:

$$Ot_{-}^{K}(\psi) = \sum_{\vec{x}\in X^K,\, acc(\vec{x})\leq\psi} \prod_{0\leq i<K} \frac{(\psi - acc(x_1,\ldots,x_i))(x_{i+1})}{L-i} \,\big|\, \vec{x}, \psi - acc(\vec{x}) \,\big\rangle.$$

2 For $\vec{x} \in X^K$ write $\varphi = acc(\vec{x})$. Then:

$$\prod_{0\leq i<K} \Big(\psi - acc(x_1,\ldots,x_i)\Big)(x_{i+1}) = \prod\nolimits_{y} \frac{\psi(y)!}{(\psi(y)-\varphi(y))!} = \frac{\psi\mathbin{!}}{(\psi-\varphi)\mathbin{!}}.$$

The right-hand side is thus independent of the sequence $\vec{x}$.

This independence means that any order of the elements of the same multiset of balls gets the same (draw) probability. This is not entirely trivial.

Proof. 1 Directly from the definition of the transition channel Ot_{-}, using Kleisli composition (3.18).

2 Write $\varphi = acc(\vec{x})$ as $\varphi = \sum_j n_j |y_j\rangle$. Then each element $y_j \in X$ occurs n_j times in the sequence $\vec{x}$. The product

$$\prod_{0\leq i<K} \Big(\psi - acc(x_1,\ldots,x_i)\Big)(x_{i+1})$$

does not depend on the order of the elements in $\vec{x}$: each element y_j occurs n_j times in this product, with multiplicities $\psi(y_j), \ldots, \psi(y_j) - n_j + 1$, independently of the exact occurrences of the y_j in $\vec{x}$. Thus:

$$\begin{aligned}
\prod_{0\leq i<K} \Big(\psi - acc(x_1,\ldots,x_i)\Big)(x_{i+1}) &= \prod\nolimits_j \psi(y_j) \cdot \ldots \cdot (\psi(y_j) - n_j + 1) \\
&= \prod\nolimits_j \psi(y_j) \cdot \ldots \cdot (\psi(y_j) - \varphi(y_j) + 1) \\
&= \prod\nolimits_j \frac{\psi(y_j)!}{(\psi(y_j)-\varphi(y_j))!} \\
&= \prod_{y\in X} \frac{\psi(y)!}{(\psi(y)-\varphi(y))!}.
\end{aligned}$$

We can extend the product over j to a product over all $y \in X$ since if $y \notin supp(\varphi)$, then, even if $\psi(y) = 0$,

$$\frac{\psi(y)!}{(\psi(y)-\varphi(y))!} = \frac{\psi(y)!}{\psi(y)!} = 1. \qquad \blacksquare$$

Theorem 3.5.4. *Consider the ordered-transition-with-deletion channel Ot_{-} on $\psi \in \mathcal{N}[L](X)$.*

1 For $K \leq L$,

$$Ot_-^K(\psi) = \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{\binom{\psi - \varphi}{\psi}}{} \big| \vec{x}, \psi - \varphi \big\rangle.$$

2 The associated K-draw channel $O_-[K] := \pi_1 \circ Ot_-^K \colon \mathcal{N}[L](X) \nrightarrow X^K$ satisfies:

$$O_-[K](\psi) = \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{\binom{\psi - \varphi}{\psi}}{} \big| \vec{x} \big\rangle = \sum_{\vec{x} \in X^K,\, \mathrm{acc}(\vec{x}) \leq \psi} \frac{\binom{\psi - \mathrm{acc}(\vec{x})}{\psi}}{} \big| \vec{x} \big\rangle.$$

Proof. 1 By combining the two items of Lemma 3.5.3 and using:

$$\prod_{0 \leq i < K} (L - i) = L \cdot (L-1) \cdot \ldots \cdot (L - K + 1) = \frac{L!}{(L-K)!},$$

we get:

$$\begin{aligned} Ot_-^K(\psi) &= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{(L-K)!}{L!} \cdot \prod\nolimits_y \frac{\psi(y)!}{(\psi(y) - \varphi(y))!} \big| \vec{x}, \psi - \varphi \big\rangle \\ &= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{(L-K)!}{\prod_y (\psi(y) - \varphi(y))!} \cdot \frac{\prod_y \psi(y)!}{L!} \big| \vec{x}, \psi - \varphi \big\rangle \\ &= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{(L-K)!}{(\psi - \varphi)!} \cdot \frac{\psi!}{L!} \big| \vec{x}, \psi - \varphi \big\rangle \\ &= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{\binom{\psi - \varphi}{\psi}}{} \big| \vec{x}, \psi - \varphi \big\rangle. \end{aligned}$$

2 Directly by the previous item. ❑

We now move from deletion to addition, that is, from Ot_- to Ot_+, still in the ordered case. The analysis is very much as in Lemma 3.5.3.

Lemma 3.5.5. *Let $\psi \in \mathcal{N}_*(X)$ be a non-empty multiset.*

1 Iterating K times gives:

$$Ot_+^K(\psi) = \sum_{\vec{x} \in X^K,\, \mathrm{acc}(\vec{x}) \leq \psi} \prod_{0 \leq i < K} \frac{(\psi + \mathrm{acc}(x_1, \ldots, x_i))(x_{i+1})}{\|\psi\| + i} \big| \vec{x}, \psi + \mathrm{acc}(\vec{x}) \big\rangle.$$

2 For $\vec{x} \in X^K$ with $\varphi = \mathrm{acc}(\vec{x})$,

$$\prod_{0 \leq i < K} \Big(\psi + \mathrm{acc}(x_1, \ldots, x_i)\Big)(x_{i+1}) = \prod_{y \in \mathrm{supp}(\psi)} \frac{(\psi(y) + \varphi(y) - 1)!}{(\psi(y) - 1)!}.$$

This expression on the right is independent of the sequence $\vec{x}$.

Proof. The first item is easy, so we concentrate on the second, in line with the proof of Lemma 3.5.3. If $\varphi = acc(\vec{x}) = \sum_j n_j|y_j\rangle$ we now have:

$$\begin{aligned}\prod_{0\leq i<K}\Big(\psi + acc(x_1,\ldots,x_i)\Big)(x_{i+1}) &= \prod_j \psi(y_j)\cdot\ldots\cdot(\psi(y_j)+\varphi(y_j)-1)\\ &= \prod_j \frac{(\psi(y_j)+\varphi(y_j)-1)!}{(\psi(y_j)-1)!}\\ &= \prod_{y\in supp(\psi)} \frac{(\psi(y)+\varphi(y)-1)!}{(\psi(y)-1)!}.\end{aligned}$$

❑

Theorem 3.5.6. *Let $\psi \in \mathcal{N}_*(X)$ be a non-empty multiset and let $K \in \mathbb{N}$. Then:*

$$Ot^K_+(\psi) = \sum_{\varphi\in\mathcal{N}[K](supp(\psi))}\ \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{\binom{}{\varphi}}\cdot\frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{\|\psi\|}{K}\!\right)}\,\big|\vec{x},\psi+\varphi\rangle.$$

The associated draw channel $O_+[K] \coloneqq \pi_1 \circ Ot^K_+ \colon \mathcal{N}_*(X) \nrightarrow \mathcal{N}[K](X)$ then is:

$$O_+[K](\psi) = \sum_{\varphi\in\mathcal{N}[K](supp(\psi))}\ \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{\binom{}{\varphi}}\cdot\frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{\|\psi\|}{K}\!\right)}\,\big|\vec{x}\rangle.$$

We recall from Definition 1.7.1 (2) that:

$$\left(\!\binom{\psi}{\varphi}\!\right) = \prod_{x\in supp(\psi)}\left(\!\binom{\psi(x)}{\varphi(x)}\!\right) = \prod_{x\in supp(\psi)}\binom{\psi(x)+\varphi(x)-1}{\varphi(x)}.$$

The multichoose coefficients $\left(\!\binom{\psi}{\varphi}\!\right)$ sum to $\left(\!\binom{\|\psi\|}{K}\!\right)$ for all φ with $\|\varphi\| = K$, see Proposition 1.7.3.

Proof. Write $L = \|\psi\| > 0$. We first note that:

$$\prod_{0\leq i<K}(L+i) = L\cdot(L+1)\cdot\ldots\cdot(L+K-1) = \frac{(L+K-1)!}{(L-1)!}.$$

In combination with Lemma 3.5.5 we get:

$$\begin{aligned}
&Ot_+^K(\psi) \\
&= \sum_{\varphi\in\mathcal{N}[K](supp(\psi))} \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{(\|\psi\|-1)!}{(\|\psi\|+K-1)!} \cdot \prod_{y\in supp(\psi)} \frac{(\psi(y)+\varphi(y)-1)!}{(\psi(y)-1)!} \,\big|\vec{x}, \psi+\varphi\big\rangle \\
&= \sum_{\varphi\in\mathcal{N}[K](supp(\psi))} \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{\prod_{y\in supp(\psi)} \frac{(\psi(y)+\varphi(y)-1)!}{(\psi(y)-1)!}}{\frac{(\|\psi\|+K-1)!}{(\|\psi\|-1)!}} \,\big|\vec{x}, \psi+\varphi\big\rangle \\
&= \sum_{\varphi\in\mathcal{N}[K](supp(\psi))} \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{\prod_y \varphi(y)!}{K!} \cdot \frac{\prod_{y\in supp(\psi)} \frac{(\psi(y)+\varphi(y)-1)!}{\varphi(y)!\cdot(\psi(y)-1)!}}{\frac{(\|\psi\|+K-1)!}{K!\cdot(\|\psi\|-1)!}} \,\big|\vec{x}, \psi+\varphi\big\rangle \\
&= \sum_{\varphi\in\mathcal{N}[K](supp(\psi))} \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{\binom{\varphi}{}} \cdot \frac{\binom{\binom{\psi}{\varphi}}{}}{\binom{\binom{\|\psi\|}{K}}{}} \,\big|\vec{x}, \psi+\varphi\big\rangle.
\end{aligned}$$

The formula for $Ot_+[K](\psi)$ is then an easy consequence. ❑

3.5.2 Unordered draws from an urn

We now concentrate on the transition channels on the right in Figure 3.1, for unordered draws. Notice that we are now using $M = \mathcal{N}(X)$ as monoid in the setting of Lemma 3.5.1. We now combine all the mathematical 'heavy lifting' in one preparatory lemma.

Lemma 3.5.7. *1 For $\omega \in \mathcal{D}(X)$ and $K \in \mathbb{N}$,*

$$Ut_0^K(\omega) = \sum_{\varphi\in\mathcal{N}[K](X)} \binom{\varphi}{} \cdot \prod\nolimits_x \omega(x)^{\varphi(x)} \,\big|\varphi, \omega\big\rangle.$$

2 For $\psi \in \mathcal{N}[L+K](X)$,

$$Ut_-^K(\psi) = \sum_{\varphi\leq_K\psi} \frac{\prod_x \binom{\psi(x)}{\varphi(x)}}{\binom{L+K}{K}} \,\big|\varphi, \psi-\varphi\big\rangle = \sum_{\varphi\leq_K\psi} \frac{\binom{\psi}{\varphi}}{\binom{\|\psi\|}{K}} \,\big|\varphi, \psi-\varphi\big\rangle.$$

3 For $\psi \in \mathcal{N}_*(X)$,

$$Ut_+^K(\psi) = \sum_{\varphi\in\mathcal{N}[K](X)} \frac{\prod_{x\in supp(\psi)} \binom{\binom{\psi(x)}{\varphi(x)}}{}}{\binom{\binom{\|\psi\|}{K}}{}} \,\big|\varphi, \psi+\varphi\big\rangle = \sum_{\varphi\in\mathcal{N}[K](X)} \frac{\binom{\binom{\psi}{\varphi}}{}}{\binom{\binom{\|\psi\|}{K}}{}} \,\big|\varphi, \psi+\varphi\big\rangle.$$

Proof. 1 We use induction on $K \in \mathbb{N}$. For $K = 0$ we have $\mathcal{N}[K](X) = \{\mathbf{0}\}$ and so:

$$Ut_0^0(\omega) = unit(\omega) = 1|\mathbf{0}, \omega\rangle = \sum_{\varphi\in\mathcal{N}[0](X)} \binom{\varphi}{} \cdot \prod\nolimits_x \omega(x)^{\varphi(x)} \,\big|\varphi, \omega\big\rangle.$$

For the induction step:

$$\begin{aligned}
Ut_0^{K+1}(\omega) &= (Ut_0^K \circ Ut_0)(\omega) \\
&\overset{(3.18)}{=} \sum_{\psi\in\mathcal{N}[1](X),\,\varphi\in\mathcal{N}[K](X)} Ut_0^K(\omega)(\varphi,\omega)\cdot Ut_0(\omega)(\psi,\omega)\,\big|\psi+\varphi,\omega\big\rangle \\
&\overset{(\mathrm{IH})}{=} \sum_{y\in X,\,\varphi\in\mathcal{N}[K](X)} \binom{\varphi}{} \cdot \Big(\prod\nolimits_x \omega(x)^{\varphi(x)}\Big)\cdot\omega(y)\,\big|1|y\rangle+\varphi,\omega\big\rangle \\
&= \sum_{\psi\in\mathcal{N}[K+1](X)} \Big(\textstyle\sum_y \binom{\psi-1|y\rangle}{}\Big)\cdot\prod\nolimits_x \omega(x)^{\psi(x)}\,\big|\psi,\omega\big\rangle \\
&\overset{(1.25)}{=} \sum_{\psi\in\mathcal{N}[K+1](X)} \binom{\psi}{}\cdot\prod\nolimits_x \omega(x)^{\psi(x)}\,\big|\psi,\omega\big\rangle.
\end{aligned}$$

2 For $K = 0$ both sides are equal $1\big|\mathbf{0},\psi\big\rangle$. Next, for a multiset $\psi\in\mathcal{N}[L+K+1](X)$ we have:

$$\begin{aligned}
Ut_-^{K+1}(\psi) &= (Ut_-^K \circ Ut_-)(\psi) \\
&\overset{(3.18)}{=} \sum_{\substack{y\in supp(\psi),\,\chi\in\mathcal{N}[L](X),\\ \varphi\le_K\psi-1|y\rangle}} Ut_-^K(\psi-1|y\rangle)(\varphi,\chi)\cdot\frac{\psi(y)}{L+K+1}\,\big|\varphi+1|y\rangle,\chi\big\rangle \\
&\overset{(\mathrm{IH})}{=} \sum_{\substack{y\in supp(\psi),\\ \varphi\le_K\psi-1|y\rangle}} \frac{\binom{\psi-1|y\rangle}{\varphi}}{\binom{L+K}{K}}\cdot\frac{\psi(y)}{L+K+1}\,\big|\varphi+1|y\rangle,\psi-1|y\rangle-\varphi\big\rangle \\
&= \sum_{\substack{y\in supp(\psi),\\ \varphi\le_K\psi-1|y\rangle}} \frac{\varphi(y)+1}{K+1}\cdot\frac{\binom{\psi}{\varphi+1|y\rangle}}{\binom{L+K+1}{K+1}}\,\big|\varphi+1|y\rangle,\psi-(\varphi+1|y\rangle)\big\rangle \\
&\qquad\text{by Exercise 1.7.6} \\
&= \sum_{\chi\le_{K+1}\psi} \frac{\chi(y)}{K+1}\cdot\frac{\binom{\psi}{\chi}}{\binom{L+K+1}{K+1}}\,\big|\chi,\psi-\chi\big\rangle \\
&= \sum_{\chi\le_{K+1}\psi} \frac{\binom{\psi}{\chi}}{\binom{L+K+1}{K+1}}\,\big|\chi,\psi-\chi\big\rangle.
\end{aligned}$$

3 The case $K = 0$ so we immediately look at, the induction step, for urn ψ

with $\|\psi\| = L > 0$.

$$
\begin{aligned}
Ut_+^{K+1}(\psi) &= (Ut_+^K \circ Ut_+)(\psi) \\
&\overset{(3.18)}{=} \sum_{y \in supp(\psi),\, \chi} Ut_+^K(\psi + 1|y\rangle)(\varphi,\chi) \cdot \frac{\psi(y)}{L} \left|\varphi + 1|y\rangle, \chi\right\rangle \\
&\overset{(IH)}{=} \sum_{y \in supp(\psi),\, \varphi \in \mathcal{N}[K](X)} \frac{\left(\!\binom{\psi+1|y\rangle}{\varphi}\!\right)}{\left(\!\binom{L+1}{K}\!\right)} \cdot \frac{\psi(y)}{L} \left|\varphi + 1|y\rangle, \psi + 1|y\rangle + \varphi\right\rangle \\
&= \sum_{y \in supp(\psi),\, \varphi \in \mathcal{N}[K](X)} \frac{\varphi(y)+1}{K+1} \cdot \frac{\left(\!\binom{\psi}{\varphi+1|y\rangle}\!\right)}{\left(\!\binom{L}{K+1}\!\right)} \left|\varphi + 1|y\rangle, \psi + (\varphi + 1|y\rangle)\right\rangle \\
&\qquad \text{by Exercise 1.7.6} \\
&= \sum_{y,\chi \in \mathcal{N}[K+1](X)} \frac{\chi(y)}{K+1} \cdot \frac{\left(\!\binom{\psi}{\chi}\!\right)}{\left(\!\binom{L}{K+1}\!\right)} \left|\chi, \psi + \chi\right\rangle \\
&= \sum_{\chi \in \mathcal{N}[K+1](X)} \frac{\left(\!\binom{\psi}{\chi}\!\right)}{\left(\!\binom{L}{K+1}\!\right)} \left|\chi, \psi + \chi\right\rangle. \qquad \blacksquare
\end{aligned}
$$

We are now in a position to describe the multinomial, hypergeometric and Pólya distributions, using iterations of the Ut_0, Ut_- and Ut_+ transition maps, followed by marginalisation.

Theorem 3.5.8. *1 The K-draw multinomial is the first marginal of K iterations of the unordered-with-replacement transition:*

$$mn[K] = \pi_1 \circ Ut_0^K =: U_0[K].$$

2 Similarly the hypergeometric distribution arises from iterated unordered-with-deletion transitions:

$$hg[K] = \pi_1 \circ Ut_-^K =: U_-[K].$$

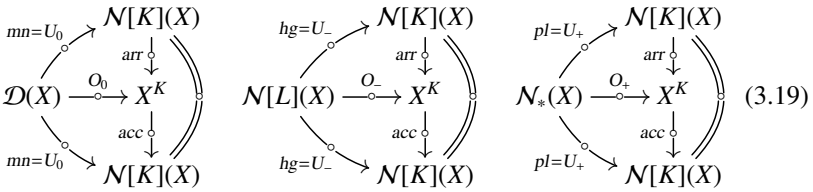
3 Finally, the Pólya distribution comes from iterated unordered-with-addition transitions, as:

$$pl[K](\psi) = \pi_1 \circ Ut_+^K =: U_+[K].$$

Proof. Directly by Lemma 3.5.7. $\blacksquare$

Together, Theorems 3.5.2, 3.5.4, 3.5.6 and 3.5.8 give precise descriptions of the six channels, for ordrered / unordered drawing with replacement / deletion / addition, as originally introduced in Table (3.3), in the introduction to this chapter. What remains to do is show that the diagrams (3.4) mentioned there commute — which relate the various forms of drawing via accumulation and

arrangement. We repeat them below for convenience, but now enriched with multinomial, hypergeometric, and Pólya channels.



(3.19)

Theorem 3.3.1 precisely says that the diagram on the left commutes. For the two other diagram we still have to do some work. This provides new characterisations of unordered draws with deletion / addition, namely as hypergeometric / Pólya followed by arrangement.

Proposition 3.5.9. *1 Accumulating ordered draws with deletion / addition yields unordered draws with deletion:*

$$(acc \otimes id) \circ Ot^K_- \;=\; Ut^K_- \qquad (acc \otimes id) \circ Ot^K_+ \;=\; Ut^K_+.$$

2 Permuting ordered draws with deletion / addition has no effect: the permutation channel $\mathsf{perm} = arr \circ acc \colon X^K \nrightarrow X^K$ *satisfies:*

$$(perm \otimes id) \circ Ot^K_- \;=\; Ot^K_- \qquad (perm \otimes id) \circ Ot^K_+ \;=\; Ot^K_+.$$

3 The diagrams in the middle and on the right in (3.19) *commute.*

Proof. 1 By combinining Theorem 3.5.4 (1) and Lemma 3.5.7 (2):

$$\begin{aligned}
((acc \otimes id) \circ Ot^K_-)(\psi) \;&=\; (acc \otimes id) \gg\!= \left(\sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{\binom{\psi - \varphi}}{\binom{\psi}} \,\middle|\, \vec{x}, \psi - \varphi \right\rangle \\
&= \sum_{\varphi \leq_K \psi} \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{\binom{\psi - \varphi}}{\binom{\psi}} \,\middle|\, acc(\vec{x}), \psi - \varphi \right\rangle. \\
&= \sum_{\varphi \leq_K \psi} \frac{\binom{\varphi} \cdot \binom{\psi - \varphi}}{\binom{\psi}} \,\middle|\, \varphi, \psi - \varphi \right\rangle. \\
&= \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L}{K}} \,\middle|\, \varphi, \psi - \varphi \right\rangle \qquad \text{by Lemma 1.6.4 (1)} \\
&= Ut^K_-(\psi).
\end{aligned}$$

Similarly, by Theorem 3.5.6,

$$\begin{aligned}((acc \otimes id) \circ Ot_+^K)(\psi) &= \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \cdot \frac{\binom{\psi}{\varphi}}{\binom{\|\psi\|}{K}} \,|\, acc(\vec{x}), \psi + \varphi \rangle. \\ &= \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} (\varphi) \cdot \frac{1}{(\varphi)} \cdot \frac{\binom{\psi}{\varphi}}{\binom{\|\psi\|}{K}} \,|\, \varphi, \psi + \varphi \rangle. \\ &= Ut_+^K(\psi).\end{aligned}$$

2 The probability for a K-tuple $\vec{x}$ in Ot_-^K Theorem 3.5.4 depends only on $acc(\vec{x})$. Hence it does not change if $\vec{x}$ is permuted. The same holds for Ot_+^K Theorem 3.5.6.

3 By the above first item we get:

$$\begin{aligned} acc \circ O_- &= acc \circ \pi_1 \circ Ot_-^K \\ &= \pi_1 \circ (acc \otimes id) \circ Ot_-^K \\ &= \pi_1 \circ Ut_-^K && \text{as shown in the first item} \\ &= hg[K] && \text{by Theorem 3.5.8 (2).} \end{aligned}$$

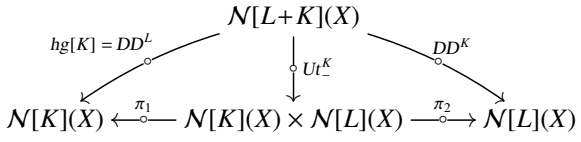
But then:

$$\begin{aligned} arr \circ hg[K] &= arr \circ acc \circ \pi_1 \circ Ut_-^K && \text{as just shown} \\ &= perm \circ \pi_1 \circ Ot_-^K \\ &= \pi_1 \circ (perm \otimes id) \circ Ot_-^K \\ &= \pi_1 \circ Ot_-^K && \text{see item (2)} \\ &= O_-. \end{aligned}$$

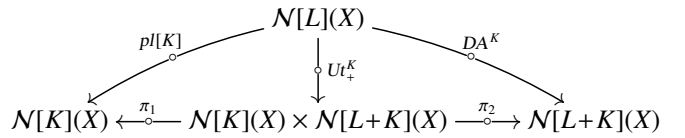
The same argument works with O_+ instead of O_-. ❑

We conclude with one more result that clarifies the different roles that the draw-and-delete DD and draw-and-add DA channels play for hypergeometric and Pólya distributions. So far we have looked only at the first marginal after iteration. including also the second marginal gives a fuller picture.

Theorem 3.5.10. *1 In the hypergeometric case, both the first and the second marginal after iterating Ut_- can be expressed in terms of the draw-and-delete map, as in:*



2 In the Pólya case, (only) the second marginal is can be expressed via the draw-and-add map:



Proof. 1 The triangle on the left commutes by Theorem 3.5.8 (2) and Theorem 3.4.1. The one on the right follows from the equation:

$$DD^K(\psi) = \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{\|\psi\|}{K}} \big|\psi - \varphi\big\rangle. \tag{3.20}$$

The proof is left as exercise to the reader.

2 Commutation of the triangle on the left holds by Theorem 3.5.8 (3). The triangle on the right follows from:

$$DA^K(\psi) = \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{\|\psi\|}{K}\!\right)} \big|\psi + \varphi\big\rangle. \tag{3.21}$$

This one is also left as exercise. ❑

Exercises

3.5.1 Consider the multiset/urn $\psi = 4|a\rangle + 2|b\rangle$ of size 6 over $\{a, b\}$. Compute $O_{-}[3](\psi) \in \mathcal{D}(\{a, b\}^3)$, both by hand and via the O_{-}-formula in Theorem 3.5.4 (2)

3.5.2 Check that:

$$\begin{aligned} &pl[2](1|a\rangle + 2|b\rangle + 1|c\rangle) \\ &= \tfrac{1}{10}\big|2|a\rangle\big\rangle + \tfrac{1}{5}\big|1|a\rangle + 1|b\rangle\big\rangle + \tfrac{3}{10}\big|2|b\rangle\big\rangle \\ &\quad + \tfrac{1}{10}\big|1|a\rangle + 1|c\rangle\big\rangle + \tfrac{1}{5}\big|1|b\rangle + 1|c\rangle\big\rangle + \tfrac{1}{10}\big|2|c\rangle\big\rangle. \end{aligned}$$

3.5.3 Show that $pl[K] \colon \mathcal{N}_*(X) \to \mathcal{D}(\mathcal{N}[K](X))$ is natural: for each function $f \colon X \to Y$ one has:

$$\begin{array}{ccc} \mathcal{N}_*(X) & \xrightarrow{pl[K]} & \mathcal{D}(\mathcal{N}[K](X)) \\ {\scriptstyle \mathcal{N}_*(f)}\downarrow & & \downarrow{\scriptstyle \mathcal{D}(\mathcal{N}[K](f))} \\ \mathcal{N}_*(Y) & \xrightarrow{pl[K]} & \mathcal{D}(\mathcal{N}[K](Y)) \end{array}$$

Hint: It suffices to show that the function Ut_+ in Figure (3.1) is natural.

3.5.4 Prove Equations (3.20) and (3.21).

3.5.5 Consider Theorem 3.5.10.

1 Show that the following diagram commutes, involving subtraction of multisets.

$$\begin{array}{ccc} \mathcal{N}[L+K](X) & \xrightarrow{\langle id\,,hg[K]\rangle} & \mathcal{N}[L+K](X) \times \mathcal{N}[K](X) \\ & \searrow^{DD^K} & \downarrow{\scriptstyle -} \\ & & \mathcal{N}[L](X) \end{array}$$

2 Idem for:

$$\begin{array}{ccc} \mathcal{N}[L](X) & \xrightarrow{\langle id\,,pl[K]\rangle} & \mathcal{N}[L](X) \times \mathcal{N}[K](X) \\ & \searrow^{DA^K} & \downarrow{\scriptstyle +} \\ & & \mathcal{N}[L+K](X) \end{array}$$

3.5.6 This exercise fills in the details of Lemma 3.5.1 and describes the relevant monad, slightly more generally, for a general monoid M, and not only for the monoid of multisets.

For an arbitrary set X define the flatten map:

$$\begin{array}{rcl} \mathcal{D}\Big(M \times \mathcal{D}(M \times X)\Big) & \xrightarrow{\mathit{flat}} & \mathcal{D}(M \times X) \\ \Omega & \longmapsto & \displaystyle\sum_{m,m',x} \Big(\textstyle\sum_\omega \Omega(m,\omega) \cdot \omega(m',x)\Big)\big|m+m',x\big\rangle. \end{array}$$

Recall that we additionally have $\mathit{unit}\colon X \to \mathcal{D}(M \times X)$ with $\mathit{unit}(x) = 1\big|0,x\big\rangle$.

1 Show that *unit* and *flat* are both natural transformations.
2 Prove the monad equations from (1.41) for $T(X) = \mathcal{D}(M \times X)$.
3 Check that the induced Kleisli composition $\circledcirc$ is the one given in Lemma 3.5.1: for $f\colon A \to \mathcal{D}(M \times B)$ and $g\colon B \to \mathcal{D}(M \times C)$,

$$\begin{array}{rcl} (g \circledcirc f)(a) & = & (\mathit{flat} \circ T(g) \circ f)(a) \\ & = & \displaystyle\sum_{m,m',c} \Big(\textstyle\sum_b f(a)(\varphi,b) \cdot g(b)(\psi,c)\Big)\big|m+m',c\big\rangle. \end{array}$$

3.6 The parallel multinomial law: four definitions

We have already seen the close connection between multisets and distributions. This section focuses on a very special 'distributivity' relation between them. It shows how a (natural) multiset of distributions can be transformed into a

distribution over multisets. This is a rather complicated operation, but it is a fundamental one. It can be described via a tensor product $\otimes$ of multinomials, and will therefore be called the *parallel multinonial law*, abbreviated as *pml*.

This law *pml* is an instance of what is called a *distributive law* in category theory. It has popped up, without explicit description, in [99, 31] and in [34], for continuous probability, to describe 'point processes' as distributions over multisets. The explicit descriptions of the law that we use below come from [78]. This law satisfies several elementary properties that combine basic elements of probability theory.

The parallel multinonial law *pml* that we are after has the following type. For a number $K \in \mathbb{N}$ and a set X it is a function:

$$\mathcal{N}[K]\big(\mathcal{D}(X)\big) \xrightarrow{\mathit{pml}[K]} \mathcal{D}\big(\mathcal{N}[K](X)\big). \tag{3.22}$$

The dependence of *pml* on K (and X) is often left implicit. Notice that *pml* can also be written as channel $\mathcal{N}[K](\mathcal{D}(X)) \nrightarrow \mathcal{N}[K](X)$. We shall frequently encounter it in this form in commuting diagrams.

This map *pml* turns a K-element multiset of distributions over X into a distribution over K-element multisets over X. It is not immediately clear how to do this. It turns out that there are several ways to describe *pml*. This section is solely devoted to defining this law, in four different manners — yielding each time the same result. The subsequent section collects basic properties of *pml*.

First definition

Since the law (3.22) is rather complicated, we start with an example.

Example 3.6.1. Let $X = \{a, b\}$ be a sample space with two distributions $\omega, \rho \in \mathcal{D}(X)$, given by:

$$\omega = \tfrac{1}{3}|a\rangle + \tfrac{2}{3}|b\rangle \qquad \text{and} \qquad \rho = \tfrac{3}{4}|a\rangle + \tfrac{1}{4}|b\rangle. \tag{3.23}$$

We will define *pml* on the multiset of distributions $2|\omega\rangle + 1|\rho\rangle$ of size $K = 3$. The result should be a distribution on multisets of size $K = 3$ over X. There are four such multisets, namely:

$$3|a\rangle \qquad 2|a\rangle + 1|b\rangle \qquad 1|a\rangle + 2|b\rangle \qquad 3|b\rangle.$$

The goal is to assign a probability to each of them. The map *pml* does this in

the following way:

$$\begin{aligned}
&pml(2|\omega\rangle + 1|\rho\rangle) \\
&= \omega(a)\cdot\omega(a)\cdot\rho(a)\big|3|a\rangle\big\rangle \\
&\quad + \Big(\omega(a)\cdot\omega(a)\cdot\rho(b) + \omega(a)\cdot\omega(b)\cdot\rho(a) + \omega(b)\cdot\omega(a)\cdot\rho(a)\Big)\big|2|a\rangle + 1|b\rangle\big\rangle \\
&\quad + \Big(\omega(a)\cdot\omega(b)\cdot\rho(b) + \omega(b)\cdot\omega(a)\cdot\rho(b) + \omega(b)\cdot\omega(b)\cdot\rho(a)\Big)\big|1|a\rangle + 2|b\rangle\big\rangle \\
&\quad + \omega(b)\cdot\omega(b)\cdot\rho(b)\big|3|b\rangle\big\rangle \\
&= \tfrac{1}{3}\cdot\tfrac{1}{3}\cdot\tfrac{3}{4}\big|3|a\rangle\big\rangle \;+\; \Big(\tfrac{1}{3}\cdot\tfrac{1}{3}\cdot\tfrac{1}{4} + \tfrac{1}{3}\cdot\tfrac{2}{3}\cdot\tfrac{3}{4} + \tfrac{2}{3}\cdot\tfrac{1}{3}\cdot\tfrac{3}{4}\Big)\big|2|a\rangle + 1|b\rangle\big\rangle \\
&\quad + \Big(\tfrac{1}{3}\cdot\tfrac{2}{3}\cdot\tfrac{1}{4} + \tfrac{2}{3}\cdot\tfrac{1}{3}\cdot\tfrac{1}{4} + \tfrac{2}{3}\cdot\tfrac{2}{3}\cdot\tfrac{3}{4}\Big)\big|1|a\rangle + 2|b\rangle\big\rangle \;+\; \tfrac{2}{3}\cdot\tfrac{2}{3}\cdot\tfrac{1}{4}\big|3|b\rangle\big\rangle \\
&= \tfrac{1}{12}\big|3|a\rangle\big\rangle \;+\; \tfrac{13}{36}\big|2|a\rangle + 1|b\rangle\big\rangle \;+\; \tfrac{4}{9}\big|1|a\rangle + 2|b\rangle\big\rangle \;+\; \tfrac{1}{9}\big|3|b\rangle\big\rangle.
\end{aligned}$$

There is a pattern. Let's try to formulate the law *pml* from (3.22) in general, for arbitrary K and X. It is defined on a natural multiset $\sum_i n_i|\omega_i\rangle$ with multiplicities $n_i \in \mathbb{N}$ satisfying $\sum_i n_i = K$, and with distributions $\omega_i \in \mathcal{D}(X)$. The number $pml(\sum_i n_i|\omega_i\rangle)(\varphi)$ describes the probability of the K-sized multiset φ over X, by using for each element occurring in φ the probability of that element in the corresponding distribution in $\sum_i n_i|\omega_i\rangle$.

In order to make this description precise we assume that the indices i are somehow ordered, say as $i_1, \ldots, i_m$ and use this ordering to form a product state:

$$\bigotimes\nolimits_i \omega_i^{n_i} = \underbrace{\omega_{i_1} \otimes \cdots \otimes \omega_{i_1}}_{n_{i_1} \text{ times}} \otimes \;\cdots\; \otimes \underbrace{\omega_{i_m} \otimes \cdots \otimes \omega_{i_m}}_{n_{i_m} \text{ times}} \in \mathcal{D}(X^K).$$

Now we formulate the first definition:

$$\begin{aligned}
pml(\textstyle\sum_i n_i|\omega_i\rangle) &\coloneqq \sum_{\vec{x}\in X^K} \Big(\bigotimes \omega_i^{n_i}\Big)(\vec{x})\,\big|\,acc(\vec{x})\big\rangle \\
&= \sum_{\varphi\in\mathcal{N}[K](X)} \left(\sum_{\vec{x}\in acc^{-1}(\varphi)} \Big(\bigotimes \omega_i^{n_i}\Big)(\vec{x})\right)\big|\varphi\big\rangle.
\end{aligned} \tag{3.24}$$

Second definition

There is an alternative formulation of the parallel multinomial law, using sums of parallel multinomial distributions via $\otimes$. This formulation is the basis for the phrase 'parallel multinomial'.

$$\begin{aligned}
pml(\textstyle\sum_i n_i|\omega_i\rangle) &\coloneqq \mathcal{D}(+)\Big(\bigotimes_i mn[n_i](\omega_i)\Big) \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \Big(\textstyle\prod_i\; mn[n_i](\omega_i)(\varphi_i)\Big)\,\big|\,\textstyle\sum_i \varphi_i\big\rangle.
\end{aligned} \tag{3.25}$$

The sum $+$ that we use here as type:

$$\mathcal{N}[n_{i_1}](X) \times \cdots \times \mathcal{N}[n_{i_m}](X) \xrightarrow{\quad + \quad} \mathcal{N}[\underbrace{n_{i_1} + \cdots + n_{i_m}}_{K}](X).$$

That, this sum has type $\prod_i \mathcal{N}[n_i](X) \to \mathcal{N}[\sum_i n_i](X)$.

This definition (3.25) may be seen as a sum of multinomials, in the style of Proposition 2.4.5. But this is justified via inclusions $\mathcal{N}[n_i](X) \hookrightarrow \mathcal{N}(X)$. This perspective is exploited in the fourth definition below.

Proposition 3.6.2. *The definitions of the law pml in* (3.24) *and* (3.25) *are equivalent.*

Proof. Because:

$$\begin{aligned}
pml(\textstyle\sum_i n_i|\omega_i\rangle) &\overset{(3.24)}{=} \sum_{\vec{x}\in X^K} \Big(\bigotimes \omega_i^{n_i}\Big)(\vec{x}) \,\Big|\, acc(\vec{x})\Big\rangle \\
&= \sum_{i,\,\vec{x}_i\in X^{n_i}} \Big(\textstyle\prod_i \omega_i^{n_i}(\vec{x}_i)\Big)\Big|\, \sum_i acc(\vec{x}_i)\Big\rangle \qquad \text{see Exercise 1.6.5} \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \Big(\textstyle\prod_i \sum_{\vec{x}_i\in acc(\varphi_i)} \omega_i^{n_i}(\vec{x}_i)\Big)\Big|\, \sum_i \varphi_i\Big\rangle \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \Big(\textstyle\prod_i \mathcal{D}(acc)(\omega_i^{n_i})(\varphi_i)\Big)\Big|\, \sum_i \varphi_i\Big\rangle \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \Big(\textstyle\prod_i (acc \circ iid(\omega_i)(\varphi_i)\Big)\Big|\, \sum_i \varphi_i\Big\rangle \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \Big(\textstyle\prod_i mn[n_i](\omega_i)(\varphi_i)\Big)\Big|\, \sum_i \varphi_i\Big\rangle.
\end{aligned}$$

This last equation is an instance of Theorem 3.3.1 (2). It leads to the second formulation of the *pml* in (3.25). ❑

Example 3.6.3. We continue Example 3.6.1 but now we describe the application of the parallel multinonial law *pml* in terms of multinomials, as in (3.25). We use the same multiset $2|\omega\rangle + 1|\rho\rangle$ of distributions ω, ρ from (3.23). The calculation of *pml* on this multiset, according to the second definition (3.25), is a bit more complicated than in Example 3.6.1 according to the first definition, since we have to evaluate the multinomial expressions. But of course the

outcome is the same.

$$\begin{aligned}
&pml(2|\omega\rangle + 1|\rho\rangle) \\
&= \sum_{\varphi\in\mathcal{N}[2](X),\,\psi\in\mathcal{N}[1](X)} mn[2](\omega)(\varphi)\cdot mn[1](\rho)(\psi)\big|\varphi+\psi\big\rangle \\
&= mn[2](\omega)(2|a\rangle)\cdot mn[1](\rho)(1|a\rangle)\big|3|a\rangle\big\rangle \\
&\quad + \Big(mn[2](\omega)(2|a\rangle)\cdot mn[1](\rho)(1|b\rangle) \\
&\qquad + mn[2](\omega)(1|a\rangle + 1|b\rangle)\cdot mn[1](\rho)(1|a\rangle)\Big)\big|2|a\rangle + 1|b\rangle\big\rangle \\
&\quad + \Big(mn[2](\omega)(1|a\rangle + 1|b\rangle)\cdot mn[1](\rho)(1|b\rangle) \\
&\qquad + mn[2](\omega)(2|b\rangle)\cdot mn[1](\rho)(1|a\rangle)\Big)\big|1|a\rangle + 2|b\rangle\big\rangle \\
&\quad + mn[2](\omega)(2|b\rangle)\cdot mn[1](\rho)(1|b\rangle)\big|3|b\rangle\big\rangle \\
&= \tbinom{2}{2,0}\cdot\omega(a)^2\cdot\tbinom{1}{1,0}\cdot\rho(a)\big|3|a\rangle\big\rangle \\
&\quad + \Big(\tbinom{2}{2,0}\cdot\omega(a)^2\cdot\tbinom{1}{1,0}\cdot\rho(b) + \tbinom{2}{1,1}\cdot\omega(a)\cdot\omega(b)\cdot\tbinom{1}{1,0}\cdot\rho(a)\Big)\big|2|a\rangle + 1|b\rangle\big\rangle \\
&\quad + \Big(\tbinom{2}{1,1}\cdot\omega(a)\cdot\omega(b)\cdot\tbinom{1}{1,0}\cdot\rho(b) + \tbinom{2}{2,0}\cdot\omega(b)^2\cdot\tbinom{1}{1,0}\cdot\rho(a)\Big)\big|1|a\rangle + 2|b\rangle\big\rangle \\
&\quad + \tbinom{2}{2,0}\cdot\omega(b)^2\cdot\tbinom{1}{1,0}\cdot\rho(b)\big|3|b\rangle\big\rangle \\
&= (\tfrac{1}{3})^2\cdot\tfrac{3}{4}\big|3|a\rangle\big\rangle \;+\; \Big((\tfrac{1}{3})^2\cdot\tfrac{1}{4} + 2\cdot\tfrac{1}{3}\cdot\tfrac{2}{3}\cdot\tfrac{3}{4}\Big)\big|2|a\rangle + 1|b\rangle\big\rangle \\
&\quad + \Big(2\cdot\tfrac{1}{3}\cdot\tfrac{2}{3}\cdot\tfrac{1}{4} + (\tfrac{2}{3})^2\cdot\tfrac{3}{4}\Big)\big|1|a\rangle + 2|b\rangle\big\rangle \;+\; (\tfrac{2}{3})^2\cdot\tfrac{1}{4}\big|3|b\rangle\big\rangle \\
&= \tfrac{1}{12}\big|3|a\rangle\big\rangle \;+\; \tfrac{13}{36}\big|2|a\rangle + 1|b\rangle\big\rangle \;+\; \tfrac{4}{9}\big|1|a\rangle + 2|b\rangle\big\rangle \;+\; \tfrac{1}{9}\big|3|b\rangle\big\rangle.
\end{aligned}$$

Indeed, this is what has been calculated in Example 3.6.1.

Third definition

Our third definition of *pml* is more abstract than the two earlier ones. It uses the coequaliser property of Proposition 3.1.1. It determines *pml* as the unique (dashed) map in:

$$\begin{array}{ccc}
\mathcal{D}(X)^K & \xrightarrow{\quad acc \quad} & \mathcal{N}[K](\mathcal{D}(X)) \\
& \searrow^{\bigotimes} & \vdots \\
& \mathcal{D}(X^K) & \vdots\ pml \\
& \searrow_{\mathcal{D}(acc)} & \downarrow \\
& & \mathcal{D}(\mathcal{N}[K](X))
\end{array} \tag{3.26}$$

There is an important side-condition in Proposition 3.1.1, namely that the composite $f := \mathcal{D}(acc)\circ\bigotimes\colon \mathcal{D}(X)^K \to \mathcal{D}(\mathcal{N}[K](X))$ is stable under permutations. This is easy: for $\omega_i \in \mathcal{D}(X)$ and $\varphi \in \mathcal{N}[K](X)$, and for a permutation

$\pi\colon \{1,\ldots,K\} \xrightarrow{\cong} \{1,\ldots,K\}$,

$$\begin{aligned} f(\omega_1,\ldots,\omega_K)(\varphi) &= \sum_{\vec{x}\in acc^{-1}(\varphi)} (\omega_1\otimes\cdots\otimes\omega_K)(x_1,\ldots,x_K) \\ &\overset{(*)}{=} \sum_{\vec{y}\in acc^{-1}(\varphi)} (\omega_1\otimes\cdots\otimes\omega_K)(y_{\pi^{-1}(1)},\ldots,y_{\pi^{-1}(K)}) \\ &= \sum_{\vec{y}\in acc^{-1}(\varphi)} (\omega_{\pi(1)}\otimes\cdots\otimes\omega_{\pi(K)})(y_1,\ldots,y_K) \\ &= f(\omega_{\pi(1)},\ldots,\omega_{\pi(K)})(\varphi). \end{aligned}$$

The marked equation $\overset{(*)}{=}$ holds because accumulation is stable under permutation. This implies that if $\vec{x}$ is in $acc^{-1}(\varphi)$, then each permutation of $\vec{x}$ is also in $acc^{-1}(\varphi)$.

Proposition 3.6.4. *The definitions of* pml *in* (3.24)*,* (3.25) *and* (3.26) *are all equivalent.*

Proof. By the uniqueness property in the triangle (3.26) it suffices to prove that *pml* as described in the first (3.24) or second (3.25) formulation makes this triangle commute. For this we use the first version. We rely again on the fact that accumulation is stable under permutation. Assume we have $\vec{\omega} = (\omega_1,\ldots,\omega_K) \in \mathcal{D}(X)^K$ with $acc(\vec{\omega}) = \sum_{i\in S} n_i|\omega_i\rangle$, for $S \subseteq \{1,\ldots,K\}$. Then:

$$\begin{aligned} (pml \circ acc)(\vec{\omega}) &= pml(\textstyle\sum_{i\in S} n_i|\omega_i\rangle) \\ &\overset{(3.24)}{=} \sum_{\vec{x}\in X^K} (\bigotimes_i \omega_i^{n_i})(\vec{x}) \,\big|\, acc(\vec{x})\big\rangle \\ &= \sum_{\vec{x}\in X^K} (\omega_1\otimes\cdots\otimes\omega_K)(\vec{x}) \,\big|\, acc(\vec{x})\big\rangle \\ &= \mathcal{D}(acc)(\textstyle\bigotimes(\vec{\omega})). \end{aligned}$$

Implicitly, for well-definedness of the first definition 3.24 of *pml* we already used that the precise ordering of states in the tensor $\bigotimes(\vec{\omega})$ is irrelevant in the formulation of *pml*. ❑

This third formulation of the parallel multinomial law is not very useful for actual calculations, like in Examples 3.6.1 and 3.6.3. But it is useful for proving properties about *pml*, via the uniqueness part of the third definition. This will be illustrated in Exercise 3.6.3 below. More generally, the fact that we have three equivalent formulations of the same law allows us to switch freely and use whichever formulation is most convenient in a particular situation.

Fourth definition

For our fourth and last definition we have to piece together some earlier observations.

1 Recall from Proposition 2.4.5 that if M is a commutative monoid, then so is the set $\mathcal{D}(M)$ of distributions on M, with sum:

$$\omega + \rho \;=\; \mathcal{D}(+)(\omega \otimes \rho) \;=\; \sum_{x_1, x_2 \in M} (\omega \otimes \rho)(x_1, x_2) \big| x_1 + x_2 \,\rangle.$$

2 Recall from Proposition 1.4.5 that such commutative monoid structure corresponds to an $\mathcal{N}$-algebra $\alpha\colon \mathcal{N}(\mathcal{D}(M)) \to \mathcal{D}(M)$, given by:

$$\begin{aligned} \alpha(\textstyle\sum_i n_i|\omega_i\,\rangle) \;&=\; \textstyle\sum_i n_i \cdot \omega_i \\ &=\; \displaystyle\sum_{\vec{x} \in M^K} (\textstyle\bigotimes_i \omega_i^{n_i})(\vec{x}) \,\big|\, \textstyle\sum \vec{x}\,\rangle \quad \text{where} \quad K = \textstyle\sum_i n_i. \end{aligned} \tag{3.27}$$

3 For an arbitrary set X, the set $\mathcal{N}(X)$ of natural multisets on X is a commutative monoid, see Lemma 1.4.2. Applying the previous two items with $M = \mathcal{N}(X)$ yields an $\mathcal{N}$-algebra:

$$\mathcal{N}\Big(\mathcal{D}\mathcal{N}(X)\Big) \longrightarrow \mathcal{D}\mathcal{N}(X) \tag{3.28}$$

It interacts with $\mathcal{N}$'s unit and flatten operations as described in Proposition 1.4.5.

We can now formulate the fourth definition:

$$\mathit{pml} \;\coloneqq\; \Big(\mathcal{N}\mathcal{D}(X) \xrightarrow{\mathcal{N}\mathcal{D}(\mathrm{unit}^{\mathcal{N}})} \mathcal{N}\mathcal{D}\mathcal{N}(X) \xrightarrow{\alpha} \mathcal{D}\mathcal{N}(X)\Big). \tag{3.29}$$

Proposition 3.6.5. *The definition of* pml *in* (3.29) *restricts to* $\mathcal{N}[K](\mathcal{D}(X)) \to \mathcal{D}(\mathcal{N}[K](X))$*, for each* $K \in \mathbb{N}$*. This restriction is the same* pml *as defined in* (3.24)*,* (3.25) *and* (3.26)*.*

Proof. We elaborate formulation (3.29), on a K-sized multiset $\sum_i n_i|\omega_i\,\rangle$.

$$\begin{aligned} \mathit{pml}(\textstyle\sum_i n_i|\omega_i\,\rangle) \;&\overset{(3.29)}{=}\; \alpha\Big(\textstyle\sum_i n_i \,\big|\, \mathcal{D}(\mathsf{unit})(\omega_i)\,\rangle\Big) \\ &\overset{(3.27)}{=}\; \sum_{\vec{\varphi} \in \mathcal{N}(X)^K} \Big(\textstyle\bigotimes_i \mathcal{D}(\mathsf{unit})(\omega_i)^{n_i}\Big)(\vec{\varphi}) \,\big|\, \textstyle\sum \vec{\varphi}\,\rangle \\ &=\; \sum_{x_1, \ldots, x_K \in X} \Big(\textstyle\bigotimes_i \omega_i^{n_i}\Big)(x_1, \ldots, x_K) \,\big|\, 1|x_1\,\rangle + \cdots + 1|x_K\,\rangle\big\rangle \\ &=\; \sum_{\vec{x} \in X^K} \Big(\textstyle\bigotimes_i \omega_i^{n_i}\Big)(\vec{x}) \,\big|\, \mathit{acc}(\vec{x})\big\rangle \end{aligned}$$

The last line coincides with the first formulation (3.24) of *pml*. ❑

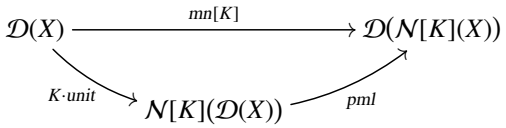
Exercises

3.6.1 Check that the multinomial channel can be obtained via the parallel multinomial law, in two different ways.

1 Use the first or second formulation, (3.24) or (3.25), of *pml* to compute that for a distribution $\omega \in \mathcal{D}(X)$ and number $K \in \mathbb{N}$ one has:

$$mn[K](\omega) \;=\; pml(K|\omega\rangle),$$

that is:



2 Prove the same thing via the third formulation (3.26), and via Theorem 3.3.1 (2).

3.6.2 Apply the multiset flatten map $flat\colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$ in the setting of Example 3.6.1 to show that:

$$flat(2|\omega\rangle + 1|\rho\rangle) \;=\; \tfrac{17}{12}|a\rangle + \tfrac{19}{12}|b\rangle \;=\; flat\Big(pml(2|\omega\rangle + 1|\rho\rangle)\Big).$$

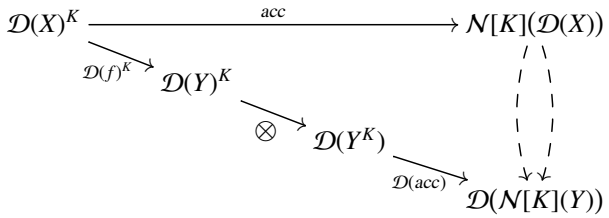
(The general formulation appears later on in Proposition 3.7.3.)

3.6.3 We claim *pml* is natural: for each function $f\colon X \to Y$ the following diagram commutes.

$$\begin{array}{ccc} \mathcal{N}[K](\mathcal{D}(X)) & \xrightarrow{pml} & \mathcal{D}(\mathcal{N}[K](X)) \\ {\scriptstyle \mathcal{N}[K](\mathcal{D}(f))}\downarrow & & \downarrow{\scriptstyle \mathcal{D}(\mathcal{N}[K](f))} \\ \mathcal{N}[K](\mathcal{D}(Y)) & \xrightarrow{pml} & \mathcal{D}(\mathcal{N}[K](Y)). \end{array}$$

1 Prove this claim via a direct calculation using the first (3.24) or second (3.25) formulation of *pml*.

2 Give an alternative proof using the uniqueness part of the third formulation (3.26), as suggested in the diagram:



3.6.4 Generalise Exercise 3.3.8 from multinomials to parallel multinomials, as in the following diagram.

$$\begin{array}{ccc} \mathcal{N}[K](\mathcal{D}(X)) \times \mathcal{N}[L](\mathcal{D}(X)) & \xrightarrow{pml \otimes pml} & \mathcal{N}[K](X) \times \mathcal{N}[L](X) \\ {\scriptstyle +}\downarrow & & \downarrow{\scriptstyle +} \\ \mathcal{N}[K+L](\mathcal{D}(X)) & \xrightarrow{\quad pml \quad} & \mathcal{N}[K+L](X) \end{array}$$

3.7 The parallel multinomial law: basic properties

This section continues the investigation of the parallel multinomial law *pml*, introduced in the previous section, in three different forms. This section focuses on some key properties of this law, including its interaction with frequentist learning, multizip and hypergeometric distributions. As we have seen, actual calculations with the parallel multinomial law *pml* quickly grow out of hand. So we might worry that proving properties also becomes tedious. But abstraction will help us. Since there is a characterisation of *pml* with a uniqueness property, in its third formulation (3.26), we can reason with the associated uniqueness proof principle. In most general form it says that $f = g$ follows from $f \circ acc = g \circ acc$, where *acc* is the acculation map, see Section 3.1, especially Proposition 3.1.1.

The next result enriches what we already now.

Proposition 3.7.1. *The parallel multinomial law* pml *is the unique channel making both rectangles below commute.*

$$\begin{array}{ccccc} \mathcal{D}(X)^K & \xrightarrow{acc} & \mathcal{N}[K](\mathcal{D}(X)) & \xrightarrow{arr} & \mathcal{D}(X)^K \\ \otimes\downarrow & & pml\downarrow & & \downarrow\otimes \\ X^K & \xrightarrow{acc} & \mathcal{N}[K](X) & \xrightarrow{arr} & X^K \end{array}$$

Proof. The rectangle on the left is the third formulation of *pml* in (3.26) and thus provides uniqueness. Commutation of the rectangle of the right follows from a uniqueness argument, using that the outer rectangle commutes by Proposition 3.1.3

$$\begin{array}{rcll} \bigotimes \circ arr \circ acc & = & arr \circ acc \circ \bigotimes & \text{by Proposition 3.1.3} \\ & = & arr \circ pml \circ acc & \text{by (3.26).} \end{array} \qquad \square$$

This result shows that *pml* is squeezed between $\bigotimes$, both on the left and on the right. We have seen in Exercise 2.4.10 that $\bigotimes$ is a distributive law. We shall prove the same about *pml* below.

But first we show how *pml* interacts with frequentist learning.

Theorem 3.7.2. *The distributive law* pml *commutes with frequentist learning, in the sense that for* $\Psi \in \mathcal{N}[K](\mathcal{D}(X))$*,*

$$\mathit{Flrn} \gg\!= \mathit{pml}(\Psi) \;=\; \mathit{flat}(\mathit{Flrn}(\Psi)).$$

Equivalently, in diagrammatic form:

$$\begin{array}{ccc} \mathcal{N}[K](\mathcal{D}(X)) & \xrightarrow{\;\mathit{pml}\;} & \mathcal{N}[K](X) \\ {\scriptstyle \mathit{Flrn}}\downarrow & & \downarrow{\scriptstyle \mathit{Flrn}} \\ \mathcal{D}(X) & \xrightarrow{\;\mathit{id}\;} & X \end{array}$$

The channel $\mathcal{D}(X) \nrightarrow X$ *at the bottom is the identity function* $\mathcal{D}(X) \to \mathcal{D}(X)$*.*

Proof. We use the second formulation of the parallel multinonial law (3.25). Let $\Psi = \sum_i n_i|\omega_i\rangle \in \mathcal{N}[K](\mathcal{D}(X))$.

$$\begin{aligned} &\mathit{Flrn} \gg\!= \mathit{pml}(\Psi) \;=\; \mathit{Flrn} \gg\!= \mathcal{D}(+)\Big(\textstyle\bigotimes_i \mathit{mn}[n_i](\omega_i)\Big) \\ &= \textstyle\sum_i \frac{n_i}{K} \cdot (\mathit{Flrn} \gg\!= \mathit{mn}[n_i](\omega_i)) \quad \text{by Exercise 3.7.2, for } K = \textstyle\sum_i n_i \\ &= \textstyle\sum_i \frac{n_i}{K} \cdot \omega_i \qquad\qquad\qquad\qquad \text{by Theorem 3.3.5} \\ &= \mathit{flat}(\textstyle\sum_i \frac{n_i}{K}|\omega_i\rangle) \\ &= \mathit{flat}(\mathit{Flrn}(\Psi)). \end{aligned}$$

❑

We include a result that generalises Exercise 3.6.2. It is a discrete version of [34, Lem. 13]. The proof below uses the 'mean' of multinomials, from Proposition 3.3.7.

Proposition 3.7.3. *Consider inclusions* $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ *and* $\mathcal{N}[K](X) \hookrightarrow \mathcal{M}(X)$*, together with the multiset flatten map* $\mathit{flat}\colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$*. Via these inclusions, one has, for* $\Psi \in \mathcal{N}[K](\mathcal{D}(X))$*,*

$$\mathit{flat}(\mathit{pml}(\Psi)) \;=\; \mathit{flat}(\Psi).$$

Proof. For $\Psi = n_1|\omega_1\rangle + \cdots + n_k|\omega_k\rangle \in \mathcal{N}[K](\mathcal{D}(X))$,

$$\begin{aligned}
\mathit{flat}(\mathit{pml}(\Psi)) &= \mathit{flat}\left(\sum_{\varphi_1 \in \mathcal{N}[n_1](X), \ldots, \varphi_k \in \mathcal{N}[n_k](X)} \left(\prod_i \mathit{mn}[n_i](\omega)(\varphi_i)\right) \middle| \sum_i \varphi_i \right\rangle \\
&= \sum_{\varphi_1 \in \mathcal{N}[n_1](X), \ldots, \varphi_k \in \mathcal{N}[n_k](X)} \left(\prod_i \mathit{mn}[n_i](\omega)(\varphi_i)\right) \cdot \left(\sum_i \varphi_i\right) \\
&= \sum_{\varphi_1 \in \mathcal{N}[n_1](X), \ldots, \varphi_k \in \mathcal{N}[n_k](X)} \left(\prod_i \mathit{mn}[n_i](\omega)(\varphi_i)\right) \cdot \varphi_1 \\
&\qquad + \cdots + \sum_{\varphi_1 \in \mathcal{N}[n_1](X), \ldots, \varphi_k \in \mathcal{N}[n_k](X)} \left(\prod_i \mathit{mn}[n_i](\omega)(\varphi_i)\right) \cdot \varphi_k \\
&= \sum_{\varphi_1 \in \mathcal{N}[n_1](X)} \mathit{mn}[n_1](\omega_1)(\varphi_1) \cdot \varphi_1 + \cdots + \sum_{\varphi_k \in \mathcal{N}[n_k](X)} \mathit{mn}[n_k](\omega_k)(\varphi_k) \cdot \varphi_k \\
&= n_1 \cdot \omega_1 + \cdots + n_k \cdot \omega_k \qquad \text{by Proposition 3.3.7} \\
&= \mathit{flat}(\Psi).
\end{aligned}$$

❑

The parallel multinomial law *pml* contains multinomial distributions. But it also interacts with the multinomial channel, as described next.

Theorem 3.7.4. *1 The parallel multinomial law* pml *commutes with multinomials in the following manner.*

$$\begin{array}{ccc}
\mathcal{D}(\mathcal{D}(X)) & \xrightarrow{\mathit{mn}[K]} & \mathcal{M}[K](\mathcal{D}(X)) \\
{\scriptstyle \mathit{flat}}\downarrow & & \downarrow{\scriptstyle \mathit{pml}} \\
\mathcal{D}(X) & \xrightarrow{\mathit{mn}[K]} & \mathcal{M}[K](X)
\end{array}$$

2 There is a second form of exchange between pml *and multinomials:*

$$\begin{array}{ccccccc}
 & {\scriptstyle \mathit{pml}}\nearrow & \mathcal{D}\mathcal{N}[K](X) & \xrightarrow{\mathit{mn}[L]} & \mathcal{D}\mathcal{N}[L]\mathcal{N}[K](X) & \searrow{\scriptstyle \mathcal{D}(\mathit{flat})} & \\
\mathcal{N}[K]\mathcal{D}(X) & & & & & & \mathcal{D}\mathcal{N}[L\cdot K](X) \\
 & {\scriptstyle \mathcal{N}[K](\mathit{mn}[L])}\searrow & \mathcal{N}[K]\mathcal{D}\mathcal{N}[L](X) & \xrightarrow[\mathit{pml}]{} & \mathcal{D}\mathcal{N}[K]\mathcal{N}[L](X) & \nearrow{\scriptstyle \mathcal{D}(\mathit{flat})} &
\end{array}$$

Proof. 1 We use that $\mathit{mn}[K] = \mathit{acc} \circ \mathit{iid}[K]$, see Theorem 3.3.1 (2), in:

$$\begin{array}{ccccc}
\mathcal{D}(\mathcal{D}(X)) & \xrightarrow{\mathit{iid}} & \mathcal{D}(X)^K & \xrightarrow{\mathit{acc}} & \mathcal{M}[K](\mathcal{D}(X)) \\
{\scriptstyle \mathit{flat}}\downarrow & & \downarrow{\scriptstyle \otimes} & & \downarrow{\scriptstyle \mathit{pml}} \\
\mathcal{D}(X) & \xrightarrow{\mathit{iid}} & X^K & \xrightarrow{\mathit{acc}} & \mathcal{M}[K](X)
\end{array}$$

The rectangle on the left commutes by Exercise 2.4.9, and the one on the right by Proposition 3.7.1.

2 We show that precomposing both legs in the diagram with the accumulation map $acc\colon \mathcal{D}(X)^K \to \mathcal{M}[K]\mathcal{D}(X)$ yields an equality. This suffices by Proposition 3.1.1.

$$\begin{aligned}
&\mathcal{D}(flat) \circ mn[L] \circ pml \circ acc \\
&\overset{(3.29)}{=} \mathcal{D}(flat) \circ mn[L] \circ \mathcal{D}(acc) \circ \bigotimes \\
&= \mathcal{D}(flat) \circ \mathcal{D}\mathcal{M}[L](acc) \circ mn[L] \circ \bigotimes \qquad \text{by naturality of } mn[L] \\
&= \mathcal{D}(flat) \circ \mathcal{D}\mathcal{M}[L](acc) \circ (mzip_K \circledcirc (mn[L] \otimes \cdots \otimes mn[L])) \\
&\qquad \text{by a generalisation of Corollary 3.3.3} \\
&= \mathcal{D}(flat) \circ \mathcal{D}\mathcal{M}[L](acc) \circ flat \circ \mathcal{D}(mzip_K) \circ \bigotimes \circ mn[L]^K \\
&= flat \circ \mathcal{D}^2(flat) \circ \mathcal{D}^2\mathcal{M}[L](acc) \circ \mathcal{D}(mzip_K) \circ \bigotimes \circ mn[L]^K \\
&= flat \circ \mathcal{D}(unit) \circ \mathcal{D}(+) \circ \bigotimes \circ mn[L]^K \qquad \text{by Proposition 3.2.5} \\
&= \mathcal{D}(+) \circ \bigotimes \circ mn[L]^K \\
&= \mathcal{D}(flat) \circ \mathcal{D}(acc) \circ \bigotimes \circ mn[L]^K \qquad \text{by Exercise 1.6.7} \\
&\overset{(3.29)}{=} \mathcal{D}(flat) \circ pml \circ acc \circ mn[L]^K \\
&= \mathcal{D}(flat) \circ pml \circ \mathcal{N}[K](mn[L]) \circ acc \qquad \text{by naturality of } acc. \qquad \blacksquare
\end{aligned}$$

We turn to *pml* and hypergeometric channels, which, as we have seen in Theorem 3.4.1, are composites of draw-and-delete maps. We know from Proposition 3.3.8 that multinomial channels commute with draw-and-delete. The same holds for the parallel multinonial law.

Proposition 3.7.5. *The following diagram commutes.*

$$\begin{array}{ccc}
\mathcal{N}[K+1](\mathcal{D}(X)) & \xrightarrow{DD} & \mathcal{N}[K](\mathcal{D}(X)) \\
{\scriptstyle pml}\downarrow & & \downarrow{\scriptstyle pml} \\
\mathcal{N}[K+1](X) & \xrightarrow{DD} & \mathcal{N}[K](X)
\end{array}$$

Proof. We use the probabilistic projection channel $ppr\colon X^{K+1} \nrightarrow X^K$ from Exercise 3.3.13 and its interaction with $\bigotimes$ in Exercise 3.3.14.

$$\begin{aligned}
DD \circledcirc pml \circledcirc acc &= DD \circledcirc acc \circ \bigotimes && \text{by (3.26)} \\
&= acc \circledcirc ppr \circledcirc \bigotimes && \text{by Exercise 3.3.13} \\
&= acc \circledcirc \bigotimes \circledcirc ppr && \text{by Exercise 3.3.14} \\
&= pml \circledcirc acc \circledcirc ppr && \text{by (3.26)} \\
&= pml \circledcirc DD \circledcirc acc && \text{by Exercise 3.3.13.} \qquad \blacksquare
\end{aligned}$$

The next consquence expresses a fundamental relationship between multinomial and hypergeometric distributions.

Corollary 3.7.6. *The parallel multinomial law commutes with the hypergeometric channel: for $L \geq K$ one has:*

$$\begin{array}{ccc} \mathcal{N}[L](\mathcal{D}(X)) & \xrightarrow{hg[K]} & \mathcal{N}[K](\mathcal{D}(X)) \\ {\scriptstyle pml}\downarrow & & \downarrow{\scriptstyle pml} \\ \mathcal{N}[L](X) & \xrightarrow{hg[K]} & \mathcal{N}[K](X) \end{array}$$

Proof. Theorem 3.4.1 shows that the hypergeometric distribution can be expressed via iterated draw-and-deletes. Hence the result follows from (iterated application of) Proposition 3.7.5. ❑

We continue to show that the parallel multinomial law *pml* commutes with the unit and flatten operations of the distribution monad. This shows that *pml* is an instance of what is called a distributive law in category theory. Such laws are important in combining different forms of computation. A notorious result, noted ome twenty years ago by Gordon Plotkin is that the powerset monad $\mathcal{P}$ does not distribute over the probability distributions monad $\mathcal{D}$. He never published this important no-go result himself. Instead, it appeared in [161, 162] (with full credits). This negative result interpreted as: there is no semantically solid way to combine non-deterministic and probabilistic computation. The fact that a distributive law for multisets and distributions does exist shows that multiset-computations and probability can be combined. Indeed, in Corollary 3.7.8 we shall see that the K-sized multiset functor $\mathcal{N}[K]$ can be 'lifted' to the category of probabilistic channels.

But first we have to show that *pml* is a distributive law. We have already seen in Exercise 3.6.3 that it is natural.

Proposition 3.7.7. *The parallel multinonial law* pml *is a* distributive law *of the K-sized multiset functor $\mathcal{N}[K]$ over the distribution monad $\mathcal{D}$. This means that* pml *commutes with the unit and flatten operations of $\mathcal{D}$, as expressed by the following two diagram.*

$$\begin{array}{ccc} \mathcal{N}[K](X) & & \\ {\scriptstyle \mathcal{N}[K](unit)}\downarrow & \searrow^{unit} & \\ \mathcal{N}[K](\mathcal{D}(X)) & \xrightarrow{pml} & \mathcal{D}(\mathcal{N}[K](X)) \end{array}$$

$$\begin{array}{ccccc} \mathcal{N}[K](\mathcal{D}^2(X)) & \xrightarrow{pml} & \mathcal{D}(\mathcal{N}[K](\mathcal{D}(X))) & \xrightarrow{\mathcal{D}(pml)} & \mathcal{D}^2(\mathcal{N}[K](X)) \\ {\scriptstyle \mathcal{N}[K](flat)}\downarrow & & & & \downarrow{\scriptstyle flat} \\ \mathcal{N}[K](\mathcal{D}(X)) & & \xrightarrow{\qquad\qquad pml \qquad\qquad} & & \mathcal{D}(\mathcal{N}[K](X)) \end{array}$$

Proof. In Exercise 2.4.10 we have seen that the big tensor $\bigotimes\colon \mathcal{D}(X)^K \to$

$\mathcal{D}(X^K)$ is a distributive law. These properties will be used to show that *pml* is a distributive law too. We exploit the uniqueness property of the third formulation 3.26.

$$\begin{array}{ll}
pml \circ \mathcal{N}[K](unit) \circ acc & \\
= pml \circ acc \circ unit^K & \text{by naturality of } acc, \text{ see Exercise 1.6.6} \\
= \mathcal{D}(acc) \circ \bigotimes \circ unit^K & \text{by (3.26)} \\
= \mathcal{D}(acc) \circ unit & \text{via the first diagram in Exercise 2.4.10} \\
= unit \circ acc & \text{by naturality of } unit.
\end{array}$$

Simarly for the flatten-diagram:

$$\begin{array}{ll}
flat \circ \mathcal{D}(pml) \circ pml \circ acc & \\
= flat \circ \mathcal{D}(pml) \circ \mathcal{D}(acc) \circ \bigotimes & \text{by (3.26)} \\
= flat \circ \mathcal{D}(\mathcal{D}(acc)) \circ \mathcal{D}(\bigotimes) \circ \bigotimes & \text{again by (3.26)} \\
= \mathcal{D}(acc) \circ flat \circ \mathcal{D}(\bigotimes) \circ \bigotimes & \text{by naturality of } flat \\
= \mathcal{D}(acc) \circ \bigotimes \circ flat^K & \text{via Exercise 2.4.10} \\
= pml \circ acc \circ flat^K & \text{once again by (3.26)} \\
= pml \circ \mathcal{N}[K](flat) \circ acc & \text{by naturality of } acc. \qquad \text{❑}
\end{array}$$

This result says that *pml* is a so-called $\mathcal{K}\ell$-law, of the functor $\mathcal{N}[K]$ over the monad $\mathcal{D}$. In general such a $\mathcal{K}\ell$-law corresponds to a lifting of the functor to the Kleisli category of channels for the monad, see [71, 84] for details.

Corollary 3.7.8. *The K-sized multiset functor $\mathcal{N}[K]\colon$ **Sets** $\to$ **Sets** lifts to a functor $\mathcal{N}[K]\colon$ **Chan**$(\mathcal{D})$ $\to$ **Chan**$(\mathcal{D})$.*

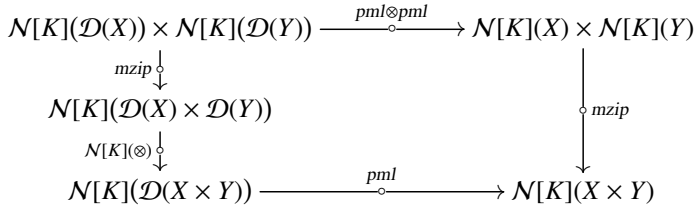
The fact that we use the same notation for two different functors may be confusing, but usually the context will tell which one is meant. These functors act in the same way on objects (sets), but differ on morphism, see at the end of the proof below.

Proof. The functor $\mathcal{N}[K]\colon$ **Chan**$(\mathcal{D})$ $\to$ **Chan**$(\mathcal{D})$ is defined on objects/sets as $X \mapsto \mathcal{N}[K](X)$. Its action on morphisms $f\colon X \nrightarrow Y$ in **Chan**$(\mathcal{D})$ is more interesting. Such an f is a function $X \to \mathcal{D}(X)$. We have to produce a function $\mathcal{N}[K](X) \to \mathcal{D}(\mathcal{N}[K](X))$. This is done via the parallel multinonial law, as composite:

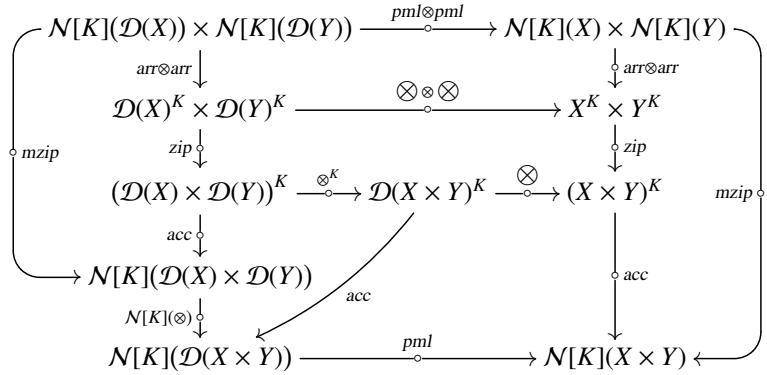
$$\mathcal{N}[K](X) \xrightarrow{\mathcal{N}[K](f)} \mathcal{N}[K](\mathcal{D}(Y)) \xrightarrow{pml} \mathcal{D}(\mathcal{N}[K](Y)). \qquad \text{❑}$$

Next we show that parallel multinomials and multizipping commute. This is a non-trivial technical result, which plays a crucial role in the subsequent result.

Lemma 3.7.9. *The following diagram commutes.*



Proof. The result follows from a big diagram chase in which the *mzip* operations on the left and on the right are expanded, according to (3.7).

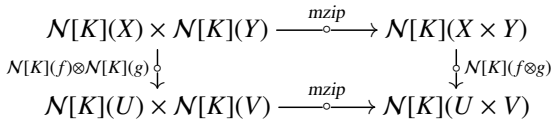


The upper rectangle commutes by Proposition 3.7.1 and the middle on by Lemma 2.4.8 (3). The lower-left subdiagram commutes by naturality of *acc* and the lower-right one via the third definition of *pml* in (3.26). ❑

Theorem 3.7.10. *The lifted functor* $\mathcal{N}[K] \colon \mathbf{Chan}(\mathcal{D}) \to \mathbf{Chan}(\mathcal{D})$ *commutes with multizipping: for channels* $f \colon X \nrightarrow U$ *and* $g \colon Y \nrightarrow V$ *one has:*

$$mzip \circ \Big(\mathcal{N}[K](f) \otimes \mathcal{N}[K](g)\Big) \;=\; \mathcal{N}[K](f \otimes g) \circ mzip. \tag{3.30}$$

Diagrammatically this amounts to:



In combination with the unit and associativity of Proposition 3.2.2 (2) *and* (5) *this means that the lifted functor* $\mathcal{N}[K] \colon \mathbf{Chan}(\mathcal{D}) \to \mathbf{Chan}(\mathcal{D})$ *is monoidal via mzip.*

Proof. This result is rather subtle, since f, g are used as channels. So when we write $\mathcal{N}[K](f)$ we mean application of the lifted functor $\mathcal{N}[K]\colon \mathbf{Chan}(\mathcal{D}) \to \mathbf{Chan}(\mathcal{D})$, as described in the last line of the proof of Corollary 3.7.8, producing another channel.

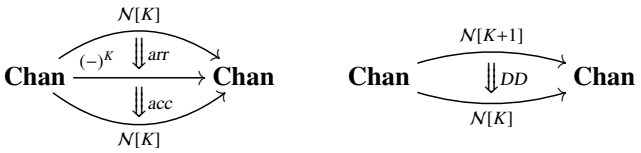
The left-hand side of the equation (3.30) thus expands as in the first equation below.

$$\begin{aligned}
&\mathit{mzip} \circ \big(\mathcal{N}[K](f) \otimes \mathcal{N}[K](g)\big) \\
&= \mathit{mzip} \circ (\mathit{pml} \otimes \mathit{pml}) \circ (\mathcal{N}[K](f) \times \mathcal{N}[K](g)) \\
&= \mathit{pml} \circ \mathcal{N}[K](\otimes) \circ \mathit{mzip} \circ (\mathcal{N}[K](f) \times \mathcal{N}[K](g)) && \text{by Lemma 3.7.9} \\
&= \mathit{pml} \circ \mathcal{N}[K](\otimes) \circ \mathcal{N}[K](f \times g) \circ \mathit{mzip} && \text{by Proposition 3.2.2 (1)} \\
&= \mathit{pml} \circ \mathcal{N}[K](f \otimes g) \circ \mathit{mzip} \\
&= \mathcal{N}[K](f \otimes g) \circ \mathit{mzip}. && \qquad \text{❑}
\end{aligned}$$

For this result we really need the multizip operation *mzip*. One may think that one can use tensors $\otimes$ instead, but the tensor-version of Lemma 3.7.9 does not hold, see Exercise 3.7.5 below.

We show that the other operations are natural with respect to channels, namely arrangement and draw-delete.

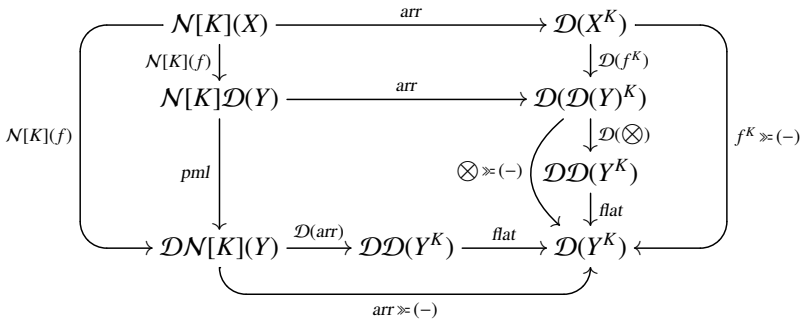
Lemma 3.7.11. *Arrangement and accumulation, and draw-delete are natural transformation in the situations:*



Proposition 3.2.2 (4) *and Exercise 3.3.15 (3) say that the natural transformations* arr *and DD are monoidal.*

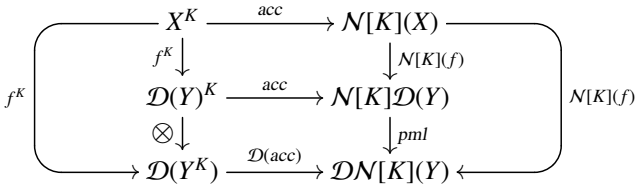
Proof. Let $f\colon X \nrightarrow Y$ be a channel. We need to prove $f^K \circ \mathit{arr} = \mathit{arr} \circ \mathcal{N}[K](f)$. We shall express this via functions as $(f^K \gg\!= (-)) \circ \mathit{arr} = (\mathit{arr} \gg\!= (-)) \circ$

$\mathcal{N}[K](f)$. This is done via the following diagram.



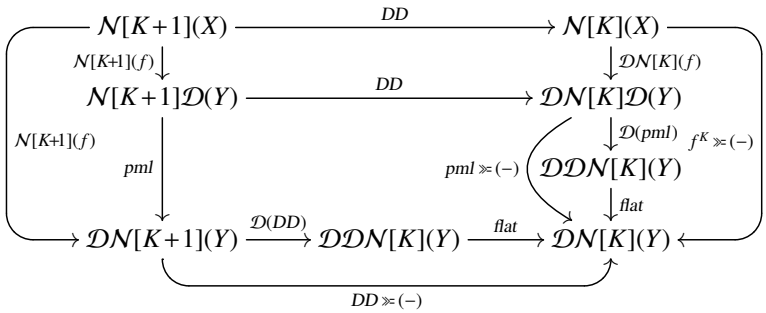
The upper rectangle is naturality of *arr*, for ordinary functions, see Exercise 2.2.7. The lower rectangle commutes by Proposition 3.7.1.

For accumulation the situation is a bit simpler. The required equality $acc \circledcirc f^K = \mathcal{N}[K](f) \circledcirc acc$ is obtained in:



The upper part is ordinary naturality of *acc* and the lower is the third formulation (3.26) of *pml*.

For naturality of draw-delete we need to prove $DD \circledcirc \mathcal{N}[K+1](f) = \mathcal{N}[K](f) \circledcirc DD$, that is $(DD \gg\!= (-)) \circ \mathcal{N}[K+1](f) = (\mathcal{N}[K](f) \gg\!= (-)) \circ DD$.



The upper diagram is naturality of draw-delete, see Exercise 2.2.8. The lower rectangle commutes by Proposition 3.7.5. ❑

We have started this section with a 3×2 table (3.3) with the six options for

drawing balls from an urn, namely orderered or unordered, and with replacement, deletion or addition. We have described these six draw maps as channels, for instance of the form $U_0 \colon \mathcal{D}(X) \nrightarrow \mathcal{N}[K](X)$ and we saw at various stages that these maps are natural in X. But this meant naturality with respect to functions. But now that we know that $\mathcal{N}[K]$ is a functor $\mathbf{Chan} \to \mathbf{Chan}$, we can also ask if these draw maps are natural with respect to channels.

This turns out to be the case. Besides $\mathcal{N}[K]$ there are other functors involved in these draw channels, namely distribution $\mathcal{D}$ and power $(-)^K$. They also have to be lifted to functors $\mathbf{Chan} \to \mathbf{Chan}$. For $\mathcal{D}$ we recall Exercise 1.9.8 which tells that $\mathcal{D}$ lifts to $\overline{\mathcal{D}} \colon \mathbf{Chan} \to \mathbf{Chan}$. The power functor $(-)^K$ lifts by using that $\bigotimes$ is a distributive law, see Exercise 2.4.10.

Theorem 3.7.12. *The following four draw maps, out of the six in Table 3.3, are natural w.r.t. channels, and thus form natural transformations in diagrams:*

K**-sized draws**	**with replacement**	**with deletion** $(L \geq K)$
ordered	$\mathbf{Chan}^{\overline{\mathcal{D}}\left(\overset{O_0}{\Longrightarrow}\right)(-)^K}\mathbf{Chan}$	$\mathbf{Chan}^{\mathcal{N}[L]\left(\overset{O_-}{\Longrightarrow}\right)(-)^K}\mathbf{Chan}$
	"independent identical"	
unordered	$\mathbf{Chan}^{\overline{\mathcal{D}}\left(\overset{U_0}{\Longrightarrow}\right)\mathcal{N}[K]}\mathbf{Chan}$	$\mathbf{Chan}^{\mathcal{N}[L]\left(\overset{U_-}{\Longrightarrow}\right)\mathcal{N}[K]}\mathbf{Chan}$
	"multinomial"	"hypergeometric"

(3.31)

These natural transformations are all monoidal, by Lemma 2.4.8 (2)*, by Corollaries 3.3.3 and 3.4.2* (7)*, and finally by Proposition 3.2.2* (4)*.*

The Pólya channel from Section 3.4 does not fit in this table since it is not natural w.r.t. channels. Intuitively, this can be explained from the fact that Pólya involves copying, and copying does not commute with channels, as we have seen early on in Exercises 2.5.10. Pólya channels are natural w.r.t. functions, see Exercise 3.5.3, and indeed, functions do commute with copying, see Exercise 2.5.12.

Proof. Naturality of $O_0 = \mathit{iid}$ in the above table (3.31) is given by the diagram on the right in Lemma 2.4.8 (1). Combining this fact with naturality of acc from Lemma 3.7.11 gives natuality for $O_- = \mathit{mn}[K]$. This uses that $\mathit{mn}[K] = \mathit{acc} \mathbin{\odot} \mathit{iid}$, see Theorem 3.3.1.

Naturality of $U_{-} = hg[K]$ follows from (channel) naturality of draw-delete in Lemma 3.7.11, since $hg[K]$ is an iteration of draw-delete's, see Theorem 3.4.1. For naturality of O_{-} we use the equation $O_{-} = arr \circ hg[K]$ from the middle diagram in (3.19), see Proposition 3.5.9 (3). Hence we are done by (channel) naturality of *arr* from Lemma 3.7.11. ❑

Exercises

3.7.1 Consider the two distributions ω, ρ in (3.23) and check yourself the following equation, which is an instance of Theorem 3.7.2.

$$\begin{aligned} Flrn \gg\!= pml(2|\omega\rangle + 1|\rho\rangle) &= \tfrac{17}{36}|a\rangle + \tfrac{19}{36}|b\rangle. \\ &= (flat \circ Flrn)(2|\omega\rangle + 1|\rho\rangle). \end{aligned}$$

3.7.2 Let $\Omega \in \mathcal{D}(\mathcal{N}[K](X))$, $\Theta \in \mathcal{D}(\mathcal{N}[L](X)$ be given. Use Exercise 2.3.3 to prove that:

$$Flrn \gg\!= \mathcal{D}(+)(\Omega \otimes \Theta) = \tfrac{K}{K+L} \cdot (Flrn \gg\!= \Omega) + \tfrac{L}{K+L} \cdot (Flrn \gg\!= \Theta).$$

3.7.3 Check that the construction of Corollary 3.7.8 indeed yields a functor $\mathcal{N}[K]\colon \mathbf{Chan}(\mathcal{D}) \to \mathbf{Chan}(\mathcal{D})$.

3.7.4 Show that the lifted functors $\mathcal{N}[K]\colon \mathcal{K}\ell(\mathcal{D}) \to \mathcal{K}\ell(\mathcal{D})$ commute with sums of multisets: for a channel $f\colon X \nrightarrow Y$,

$$\begin{array}{ccc} \mathcal{N}[K](X) \times \mathcal{N}[L](Y) & \xrightarrow{+} & \mathcal{N}[K+L](X) \\ {\scriptstyle \mathcal{N}[K](f)\otimes\mathcal{N}[L](f)}\downarrow & & \downarrow{\scriptstyle \mathcal{N}[K+L](f)} \\ \mathcal{N}[K](Y) \times \mathcal{N}[L](Y) & \xrightarrow{+} & \mathcal{N}[K+L](Y) \end{array}$$

3.7.5 The parallel multinomial law *pml* does *not* commute with tensors (of multisets and distributions), as in the following diagram.

$$\begin{array}{ccc} \mathcal{N}[K](\mathcal{D}(X)) \times \mathcal{N}[L](\mathcal{D}(Y)) & \xrightarrow{pml\otimes pml} & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \\ {\scriptstyle \otimes}\downarrow & & \downarrow \\ \mathcal{N}[K\cdot L](\mathcal{D}(X) \times \mathcal{D}(Y)) & \neq & \otimes\downarrow \\ {\scriptstyle \mathcal{N}[K\cdot L](\otimes)}\downarrow & & \downarrow \\ \mathcal{N}[K\cdot L](\mathcal{D}(X \times Y)) & \xrightarrow{pml} & \mathcal{N}[K\cdot L](X \times Y) \end{array}$$

Take for instance $X = \{a, b\}$, $Y = \{0, 1\}$ with $K = 2$, $L = 1$ with

$$\text{distributions: } \begin{cases} \omega = \frac{3}{4}|a\rangle + \frac{1}{4}|b\rangle \\ \rho = \frac{2}{3}|0\rangle + \frac{1}{3}|1\rangle \end{cases} \qquad \text{and multisets: } \begin{cases} \varphi = 2|\omega\rangle \\ \psi = 1|\rho\rangle \end{cases}$$

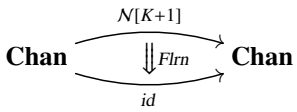
1 Calculate:

$$\begin{aligned}&(\otimes \circ (pml \otimes pml))(\varphi, \psi)\\ &= \tfrac{3}{8}\big|2|a,0\rangle\big\rangle + \tfrac{1}{4}\big|1|a,0\rangle + 1|b,0\rangle\big\rangle + \tfrac{1}{24}\big|2|b,0\rangle\big\rangle\\ &\qquad + \tfrac{3}{16}\big|2|a,1\rangle\big\rangle + \tfrac{1}{8}\big|1|a,1\rangle + 1|b,1\rangle\big\rangle + \tfrac{1}{48}\big|2|b,1\rangle\big\rangle.\end{aligned}$$

2 And also:

$$\begin{aligned}&(pml \circ \mathcal{M}[K\cdot L](\otimes) \circ \otimes)(\varphi, \psi)\\ &= \tfrac{1}{4}\big|2|a,0\rangle\big\rangle + \tfrac{1}{16}\big|2|a,1\rangle\big\rangle + \tfrac{1}{36}\big|2|b,0\rangle\big\rangle + \tfrac{1}{144}\big|2|b,1\rangle\big\rangle\\ &\qquad + \tfrac{1}{4}\big|1|a,0\rangle + 1|a,1\rangle\big\rangle + \tfrac{1}{6}\big|1|a,0\rangle + 1|b,0\rangle\big\rangle\\ &\qquad + \tfrac{1}{12}\big|1|a,0\rangle + 1|b,1\rangle\big\rangle + \tfrac{1}{12}\big|1|a,1\rangle + 1|b,0\rangle\big\rangle\\ &\qquad + \tfrac{1}{24}\big|1|a,1\rangle + 1|b,1\rangle\big\rangle + \tfrac{1}{36}\big|1|b,0\rangle + 1|b,1\rangle\big\rangle.\end{aligned}$$

3.7.6 Show that frequentist learning is a natural transformation in:



3.8 Parallel multinomials as law of monads

There is one thing we still wish to do in relation to the parallel multinomial law. We have described it as a map $pml[K]\colon \mathcal{N}[K]\mathcal{D}(X) \to \mathcal{D}\mathcal{N}[K](X)$ in a restricted manner, namely restricted to multisets of size K. What if we drop this restriction?

This question makes sense since the fourth formulation (3.29) describes *pml* simply as a map $\mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$, without size restrictions. It is in fact $pml[K]$, for the appropriate size K determined by the input. Indeed, we describe *pml* in (3.29) as:

$$pml(\varphi) = pml[\,\|\varphi\|\,](\varphi), \qquad \text{where, recall} \quad \|\varphi\| = \textstyle\sum_x \varphi(x).$$

This allows us to see that the unrestricted version $pml\colon \mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$ is natural in X, using that $\mathcal{N}$ preserves size, see Exercise 1.5.2. Thus, let $f\colon X \to Y$ and $\varphi = \sum_i n_i|\omega_i\rangle \in \mathcal{N}(\mathcal{D}(X))$ with $\|\varphi\| = \sum_i n_i = K$. Then:

$$\Big\|\, \mathcal{N}\mathcal{D}(f)(\varphi) \,\Big\| = \Big\|\, \textstyle\sum_i n_i \big|\mathcal{D}(f)(\omega_i)\big\rangle \,\Big\| = \textstyle\sum_i n_i = K.$$

Hence:

$$\begin{aligned} pml\Big(\mathcal{N}\mathcal{D}(f)(\varphi)\Big) &= pml[\,\|\mathcal{N}[K]\mathcal{D}(f)(\varphi)\|\,]\Big(\mathcal{N}[K]\mathcal{D}(f)(\varphi)\Big) \\ &= pml[K]\Big(\mathcal{N}[K]\mathcal{D}(f)(\varphi)\Big) \\ &= \mathcal{D}\mathcal{N}[K](f)\Big(pml[K](\varphi)\Big) \qquad \text{by naturality of } pml[K] \\ &= \mathcal{D}\mathcal{N}(f)\Big(pml(\varphi)\Big). \end{aligned}$$

We have seen in Proposition 3.7.7 that *pml* interacts appropriately with the unit and flatten operations of the distribution monad $\mathcal{D}$. Since we now use *pml* in unrestricted form we can also look at interaction with the unit and flatten of the (natural) multiset monad $\mathcal{N}$.

Lemma 3.8.1. *The parallel multinomial law* $\mathrm{pml}\colon \mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$ *commutes in the following way with the unit and flatten operations of the monad* $\mathcal{N}$.

$$\begin{array}{ccc} \mathcal{D}(X) & \xrightarrow{\mathcal{D}(unit)} & \\ {\scriptstyle unit}\downarrow & & \\ \mathcal{N}\mathcal{D}(X) & \xrightarrow{pml} & \mathcal{D}\mathcal{N}(X) \end{array} \qquad \begin{array}{ccccc} \mathcal{N}^2\mathcal{D}(X) & \xrightarrow{\mathcal{N}(pml)} & \mathcal{N}\mathcal{D}\mathcal{N}(X) & \xrightarrow{pml} & \mathcal{D}\mathcal{N}^2(X) \\ {\scriptstyle flat}\downarrow & & & & \downarrow{\scriptstyle \mathcal{D}(flat)} \\ \mathcal{N}\mathcal{D}(X) & & \xrightarrow{\qquad\qquad pml \qquad\qquad} & & \mathcal{D}\mathcal{N}(X) \end{array}$$

Moreover, the law $\mathrm{pml}\colon \mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$ *is a homomorphism of monoids.*

Proof. The equation for the units is easy. For $\omega \in \mathcal{D}(X)$ we have, by Exercise 3.6.1,

$$\begin{aligned} (pml \circ unit)(\omega) \;=\; pml(1|\omega\rangle) \;=\; mn[1](\omega) &= \sum_{x\in X} \omega(x)\Big|\, 1|x\rangle\Big\rangle \\ &= \mathcal{D}(unit)(\omega). \end{aligned}$$

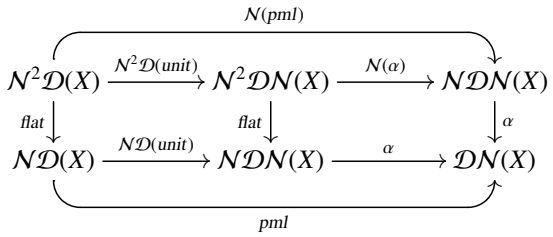
For flatten we have to do a bit more work. We recall the above $\mathcal{N}$-algebra $\alpha\colon \mathcal{N}(\mathcal{D}(M)) \to \mathcal{D}(M)$ in (3.27), for a commutative monoid M. We also recall that by Proposition 1.4.5 the following two diagrams commute, where $f\colon M_1 \to M_2$ is a map of (commutative) monoids.

$$\begin{array}{ccc} \mathcal{N}^2\mathcal{D}(M) & \xrightarrow{\mathcal{N}(\alpha)} & \mathcal{N}\mathcal{D}(M) \\ {\scriptstyle flat}\downarrow & & \downarrow{\scriptstyle \alpha} \\ \mathcal{N}\mathcal{D}(M) & \xrightarrow{\alpha} & \mathcal{D}(M) \end{array} \qquad \begin{array}{ccc} \mathcal{N}\mathcal{D}(M_1) & \xrightarrow{\mathcal{N}\mathcal{D}(f)} & \mathcal{N}\mathcal{D}(M_2) \\ {\scriptstyle \alpha}\downarrow & & \downarrow{\scriptstyle \alpha} \\ \mathcal{D}(M_1) & \xrightarrow{\mathcal{D}(f)} & \mathcal{D}(M_2) \end{array} \tag{3.32}$$

We use the fourth formulation (3.29) namely $pml = \alpha \circ \mathcal{N}(\mathcal{D}(unit))$. Then:

$$\begin{aligned}
&\mathcal{D}(flat) \circ pml \circ \mathcal{N}(pml) \\
&= \mathcal{D}(flat) \circ \alpha \circ \mathcal{N}\mathcal{D}(unit) \circ \mathcal{N}(pml) && \text{by (3.29)} \\
&= \alpha \circ \mathcal{N}\mathcal{D}(flat) \circ \mathcal{N}\mathcal{D}(unit) \circ \mathcal{N}(pml) && \text{by (3.32), on the right} \\
&= \alpha \circ \mathcal{N}(pml) && \text{by a flatten-unit law} \\
&= \alpha \circ \mathcal{N}(\alpha \circ \mathcal{N}\mathcal{D}(unit)) && \text{by (3.29) again} \\
&= \alpha \circ flat \circ \mathcal{N}\mathcal{N}\mathcal{D}(unit) && \text{by (3.32), on the left} \\
&= \alpha \circ \mathcal{N}\mathcal{D}(unit) \circ flat && \text{by naturality} \\
&= pml \circ flat && \text{once again by (3.29).}
\end{aligned}$$

In order to show that the parallel multinomial law $pml \colon \mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$ is a map of monoids it suffices by Proposition 1.4.5 to show that the diagram (1.21) commutes. This is the outer rectangle in:



The rectangle on the left commutes by naturality of *flat*; the one on the right is an instance of the rectangle on the left in (3.32). ❑

Theorem 3.8.2. *The composite* $\mathcal{D}\mathcal{N} \colon$ **Sets** $\to$ **Sets** *is a monad, with* $unit \colon X \to \mathcal{D}\mathcal{N}(X)$ *and* $flat \colon \mathcal{D}\mathcal{N}\mathcal{D}\mathcal{N}(X) \to \mathcal{D}\mathcal{N}(X)$ *operations:*

$$unit \;:=\; \left(\begin{array}{ccccc} & & \mathcal{D}(X) & & \\ & \overset{unit}{\nearrow} & & \overset{\mathcal{D}(unit)}{\searrow} & \\ X & & & & \mathcal{D}\mathcal{N}(X) \\ & \underset{unit}{\searrow} & & \underset{\mathcal{N}(unit)}{\nearrow} & \\ & & \mathcal{N}(X) & & \end{array} \right)$$

$$flat \;:=\; \left(\begin{array}{ccccccc} \mathcal{D}\mathcal{N}\mathcal{D}\mathcal{N}(X) & \xrightarrow{\mathcal{D}(pml)} & \mathcal{D}^2\mathcal{N}^2(X) & & \overset{flat}{\nearrow} & \mathcal{D}\mathcal{N}^2(X) & \overset{\mathcal{D}(flat)}{\searrow} \\ & & & \underset{\mathcal{D}^2(flat)}{\searrow} & & & \mathcal{D}\mathcal{N}(X) \\ & & & & \mathcal{D}^2\mathcal{N}(X) & \underset{flat}{\nearrow} & \end{array} \right)$$

Proof. This is a standard result in category, originally from Beck, see *e.g.* [12, 9, 71]. The diamonds in the above descriptions of unit and flatten commute by naturality. ❑

We now know that the composite functor $\mathcal{DN}\colon \mathbf{Sets} \to \mathbf{Sets}$ is a monad. Our next step is to show that there is a map of monads $\mathcal{DM} \Rightarrow \mathcal{M}$, where $\mathcal{M}\colon \mathbf{Sets} \to \mathbf{Sets}$ is the multiset monad, with non-negative numbers as multiplicities — and not just natural numbers, as for natural multisets, in $\mathcal{N}$. This map of monads is introduced in [34], under the name 'intensity'. Therefor we shall write it as $its\colon \mathcal{DM} \Rightarrow \mathcal{M}$.

There are obvious inclusions $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ and $\mathcal{N}(X) \hookrightarrow \mathcal{M}(X)$ that we used before. We now need to formalise the situation and so we shall use explicit names for these inclusions, as natural transformations, namely:

$$\mathcal{D} \Longrightarrow \mathcal{M} \qquad \mathcal{N} \Longrightarrow \mathcal{M}.$$

These σ and τ are maps of monads. They are used implicitly for instance in the equation $\mathit{flat}(\mathit{pml}(\Psi)) = \mathit{flat}(\Psi)$ in Proposition 3.7.3. As a commuting diagram, it looks as follows.

$$\begin{array}{ccccccc} \mathcal{ND} & \xrightarrow{\mathcal{N}(\sigma)} & \mathcal{NM} & \xrightarrow{\tau} & \mathcal{MM} & \xrightarrow{\mathit{flat}} & \\ {\scriptstyle \mathit{pml}}\downarrow & & & & & & \mathcal{M} \\ \mathcal{DN} & \xrightarrow{\sigma} & \mathcal{MN} & \xrightarrow{\mathcal{M}(\tau)} & \mathcal{MM} & \xrightarrow{\mathit{flat}} & \end{array} \tag{3.33}$$

Theorem 3.8.3 (From [34])**.** *The* intensity *natural transformation* $its\colon \mathcal{DN} \Rightarrow \mathcal{M}$ *has components:*

$$its \;:=\; \left(\mathcal{DN}(X) \xrightarrow{\sigma} \mathcal{MN}(X) \xrightarrow{\mathcal{M}(\tau)} \mathcal{MM}(X) \xrightarrow{\mathit{flat}} \mathcal{M}(X) \right) \tag{3.34}$$

$$\mathcal{DN}(X) \xrightarrow{\mathcal{D}(\tau)} \mathcal{DM}(X) \xrightarrow{\sigma} \mathcal{MM}(X)$$

This intensity its *is a map of monads.*

This definition 3.34 looks impressive, but for practical purposes we can just write $its(\Psi) = \mathit{flat}(\Psi)$, leaving inclusions σ, τ implicit. However, in the proof below we will be very precise and make them explicit.

Proof. Commutation of intensity with units is easy, with $\mathit{unit}^{\mathcal{DN}}$ for the monad $\mathcal{DN}$ from Theorem 3.8.2:

$$\begin{array}{rcll} its \circ \mathit{unit}^{\mathcal{DN}} & \overset{(3.34)}{=} & \mathit{flat}^{\mathcal{M}} \circ \mathcal{M}(\tau) \circ \sigma \circ \mathcal{D}(\mathit{unit}^{\mathcal{N}}) \circ \mathit{unit}^{\mathcal{D}} & \\ & = & \mathit{flat}^{\mathcal{M}} \circ \mathcal{M}(\tau) \circ \mathcal{M}(\mathit{unit}^{\mathcal{N}}) \circ \sigma \circ \mathit{unit}^{\mathcal{D}} & \\ & = & \mathit{flat}^{\mathcal{M}} \circ \mathcal{M}(\mathit{unit}^{\mathcal{M}}) \circ \mathit{unit}^{\mathcal{M}} & \text{since } \tau, \sigma \text{ are maps of monads} \\ & = & \mathit{unit}^{\mathcal{M}}. & \end{array}$$

Commutation with flatten maps is much more laborious and involves a long calculation. Figure 3.2 provides all required equational steps, using the standard equations for (maps of) monads. ❑

$$\begin{aligned}
its \circ flat^{\mathcal{DN}} &\overset{(3.34)}{=} flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\tau) \circ \mathcal{D}(flat^{\mathcal{N}}) \circ flat^{\mathcal{D}} \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{DM}(\tau) \circ \mathcal{D}(\tau) \circ flat^{\mathcal{D}} \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \mathcal{M}^2(\tau) \circ \mathcal{M}(\tau) \circ \sigma \circ flat^{\mathcal{D}} \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \mathcal{M}^2(\tau) \circ \mathcal{M}(\tau) \circ flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \mathcal{M}^3(\tau) \circ \mathcal{M}^2(\tau) \circ \sigma \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \mathcal{M}^3(\tau) \circ \sigma \circ \mathcal{DM}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \mathcal{M}^2(\tau) \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \sigma \circ \mathcal{DM}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \sigma \circ \mathcal{DM}(\tau) \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{DM}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&\overset{(3.33)}{=} flat^{\mathcal{M}} \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \sigma \circ \mathcal{DM}(\tau) \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{D}(\tau) \circ \mathcal{DN}(\sigma) \\
&= flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{DM}^2(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{DN}(\sigma) \\
&= flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{DM}(flat^{\mathcal{M}}) \circ \mathcal{DM}^2(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{DN}(\sigma) \\
&= flat^{\mathcal{M}} \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \sigma \circ \mathcal{DM}(flat^{\mathcal{M}}) \circ \mathcal{DM}^2(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{DN}(\sigma) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\tau) \circ \mathcal{DN}(flat^{\mathcal{M}}) \circ \mathcal{DNM}(\tau) \circ \mathcal{DN}(\sigma) \\
&\overset{(3.34)}{=} flat^{\mathcal{M}} \circ its \circ \mathcal{DN}(its).
\end{aligned}$$

Figure 3.2 Equational proof that the intensity natural transformation *its* from (3.34) commutes with flattens, as part of the proof of Theorem 3.8.3.

There is more to say about this situation.

Proposition 3.8.4. *For each set X, the intensity map is a homomorphism of monoids, so that we can rephrase* (3.33) *as a triangle of monoid homomorphisms:*

$$\begin{array}{ccc}
(\mathcal{ND}(X), +, 0) & \xrightarrow{\;pml\;} & (\mathcal{DN}(X), +, 0) \\
& \searrow & \downarrow{\scriptstyle its} \\
& & (\mathcal{M}(X), +, 0).
\end{array}$$

Proof. We recall that the monoid structure on $\mathcal{DN}(X)$ is defined in Proposition 2.4.5, and induces the map $\alpha\colon \mathcal{NDN}(X) \to \mathcal{DN}(X)$ used above. The zero element in $\mathcal{DN}(X)$ is $1|\mathbf{0}\rangle$, where $\mathbf{0} \in \mathcal{N}(X)$ is the empty multiset. It satisfies, for $x \in X$,

$$its(1|\mathbf{0}\rangle)(x) \;=\; flat(1|\mathbf{0}\rangle)(x) \;=\; 1 \cdot \mathbf{0}(x) \;=\; 0.$$

Hence $its(1|\mathbf{0}\rangle) = \mathbf{0}$.

For distributions $\omega, \rho \in \mathcal{DN}(X)$ we have $its(\omega + \rho) = its(\omega) + its(\rho)$ since

for each $x \in X$,

$$\begin{aligned}
its(\omega + \rho)(x) &= flat(\omega + \rho)(x) = \sum_{\varphi \in \mathcal{N}(X)} (\omega + \rho)(\varphi) \cdot \varphi(x) \\
&\overset{(2.18)}{=} \sum_{\varphi \in \mathcal{N}(X)} \mathcal{D}(+)(\omega \otimes \rho)(\varphi) \cdot \varphi(x) \\
&= \sum_{\psi, \chi \in \mathcal{N}(X)} \omega(\psi) \cdot \rho(\chi) \cdot (\psi + \chi)(x) \\
&= \sum_{\psi, \chi \in \mathcal{N}(X)} \omega(\psi) \cdot \rho(\chi) \cdot (\psi(x) + \chi(x)) \\
&= \sum_{\psi \in \mathcal{N}(X)} \omega(\psi) \cdot \left(\sum_{\chi \in \mathcal{N}(X)} \rho(\chi) \right) \cdot \psi(x) \;+\; \sum_{\chi \in \mathcal{N}(X)} \left(\sum_{\psi \in \mathcal{N}(X)} \omega(\psi) \right) \cdot \rho(\chi) \cdot \chi(x) \\
&= flat(\omega)(x) + flat(\rho)(x) \\
&= \Big(its(\omega) + its(\rho) \Big)(x).
\end{aligned}$$

❑

Exercises

3.8.1 Check that the intensity natural transformation $its \colon \mathcal{DN} \Rightarrow \mathcal{M}$ defined in (3.34) restricts to $\mathcal{DN}[K] \Rightarrow \mathcal{M}[K]$, for each $K \in \mathbb{N}$.

3.8.2 Check that the 'mean' results for multinomial, hypergeometric and Pólya distributions from Propositions 3.3.7 and 3.4.7 can be reformulated in terms of intensity as:

$$\begin{aligned}
its\Big(mn[K](\omega)\Big) &= K \cdot \omega \\
its\Big(hg[K](\psi)\Big) &= K \cdot Flrn(\psi) \\
its\Big(pl[K](\psi)\Big) &= K \cdot Flrn(\psi).
\end{aligned}$$

3.8.3 In Corollary 3.7.8 we have seen the lifted functor $\mathcal{N}[K] \colon \mathbf{Chan}(\mathcal{D}) \to \mathbf{Chan}(\mathcal{D})$. The flatten operation $flat \colon \mathcal{NN} \Rightarrow \mathcal{N}$ for (natural) multisets, from Subsection 1.4.2, restricts to $\mathcal{N}[K]\mathcal{N}[L] \Rightarrow \mathcal{N}[K \cdot L]$, making $\mathcal{N}[K] \colon \mathbf{Sets} \to \mathbf{Sets}$ into what is called a *graded* monad, see *e.g.* [123, 54]. Also the lifted functor $\mathcal{N}[K] \colon \mathbf{Chan}(\mathcal{D}) \to \mathbf{Chan}(\mathcal{D})$ is such a graded monad, essentially by Lemma 3.8.1.

The aim of this exercise is to show that these (lifted) flattens do not commute with multizip. This means that the following diagram of

channels does *not* commute.

$$\begin{array}{ccc}
\mathcal{N}[K]\mathcal{N}[L](X) \times \mathcal{N}[K]\mathcal{N}[L](Y) & \xrightarrow{\mathit{flat}\otimes\mathit{flat}} & \mathcal{N}[K \cdot L](X) \times \mathcal{N}[K \cdot L](Y) \\
\downarrow{\scriptstyle \mathit{mzip}} & & \\
\mathcal{N}[K](\mathcal{N}[L](X) \times \mathcal{N}[L](Y)) & \neq & \downarrow{\scriptstyle \mathit{mzip}} \\
\downarrow{\scriptstyle \mathcal{N}[K](\mathit{mzip})} & & \\
\mathcal{N}[K]\mathcal{N}[L](X \times Y) & \xrightarrow{\quad\mathit{flat}\quad} & \mathcal{N}[K \cdot L](X \times Y)
\end{array}$$

Elaborating a counterexample is quite initimidating, so we proceed step by step. We take as spaces:

$$X = \{a, b\} \qquad \text{and} \qquad Y = \{0, 1\}.$$

We use $K = 2$ and $L = 3$ for the multisets of multisets:

$$\begin{aligned}
\Phi &= 1\big|2|a\rangle + 1|b\rangle\big\rangle + 1\big|1|a\rangle + 2|b\rangle\big\rangle \in \mathcal{N}[2]\mathcal{N}[3](X) \\
\Psi &= 1\big|2|0\rangle + 1|1\rangle\big\rangle + 1\big|3|1\rangle\big\rangle \in \mathcal{N}[2]\mathcal{N}[3](Y).
\end{aligned}$$

1 Check that going east-south in the above diagram yields:

$$\begin{aligned}
\mathit{mzip}(\mathit{flat}(\Phi), \mathit{flat}(\Psi)) &= \mathit{mzip}(3|a\rangle + 3|b\rangle, 2|0\rangle + 4|1\rangle) \\
&= \tfrac{1}{5}\big|3|a, 1\rangle + 2|b, 0\rangle + 1|b, 1\rangle\big\rangle \\
&\quad + \tfrac{3}{5}\big|1|a, 0\rangle + 2|a, 1\rangle + 1|b, 0\rangle + 2|b, 1\rangle\big\rangle \\
&\quad + \tfrac{1}{5}\big|2|a, 0\rangle + 1|a, 1\rangle + 3|b, 1\rangle\big\rangle.
\end{aligned}$$

2 The other path, south-east, will be done in several steps. Write $\Phi = 1|\varphi_1\rangle + 1|\varphi_2\rangle$ and $\Psi = 1|\psi_1\rangle + 1|\psi_2\rangle$ where:

$$\begin{cases} \varphi_1 = 2|a\rangle + 1|b\rangle \\ \varphi_2 = 1|a\rangle + 2|b\rangle \end{cases} \qquad \begin{cases} \psi_1 = 2|0\rangle + 1|1\rangle \\ \psi_2 = 3|1\rangle. \end{cases}$$

Show then that:

$$\mathit{mzip}(\Phi, \Psi) = \tfrac{1}{2}\big|1|\varphi_1, \psi_1\rangle + 1|\varphi_2, \psi_2\rangle\big\rangle + \tfrac{1}{2}\big|1|\varphi_1, \psi_2\rangle + 1|\varphi_2, \psi_1\rangle\big\rangle.$$

3 Show next:

$$\begin{aligned}
\mathit{mzip}(\varphi_1, \psi_1) &= \tfrac{2}{3}\big|1|a, 0\rangle + 1|a, 1\rangle + 1|b, 0\rangle\big\rangle + \tfrac{1}{3}\big|2|a, 0\rangle + 1|b, 1\rangle\big\rangle \\
\mathit{mzip}(\varphi_1, \psi_2) &= 1\big|2|a, 1\rangle + 1|b, 1\rangle\big\rangle \\
\mathit{mzip}(\varphi_2, \psi_1) &= \tfrac{1}{3}\big|1|a, 1\rangle + 2|b, 0\rangle\big\rangle + \tfrac{2}{3}\big|1|a, 0\rangle + 1|b, 0\rangle + 1|b, 1\rangle\big\rangle \\
\mathit{mzip}(\varphi_2, \psi_2) &= 1\big|1|a, 1\rangle + 2|b, 1\rangle\big\rangle.
\end{aligned}$$

4 Show now that:

$$\begin{aligned}
&pml\Big(1|mzip(\varphi_1,\psi_1)\rangle + 1|mzip(\varphi_2,\psi_2)\rangle\Big)\\
&= \tfrac{2}{3}\big|1|1|a,0\rangle + 1|a,1\rangle + 1|b,0\rangle\rangle \;+\; 1|1|a,1\rangle + 2|b,1\rangle\rangle\big\rangle\\
&\quad + \tfrac{1}{3}\big|1|2|a,0\rangle + 1|b,1\rangle\rangle \;+\; 1|1|a,1\rangle + 2|b,1\rangle\rangle\big\rangle\\
&pml\Big(1|mzip(\varphi_1,\psi_2)\rangle + 1|mzip(\varphi_2,\psi_1)\rangle\Big)\\
&= \tfrac{1}{3}|1|1|a,1\rangle + 2|b,0\rangle\rangle \;+\; 1|2|a,1\rangle + 1|b,1\rangle\rangle\rangle +\\
&\quad + \tfrac{2}{3}|1|1|a,0\rangle + 1|b,0\rangle + 1|b,1\rangle\rangle \;+\; 1|2|a,1\rangle + 1|b,1\rangle\rangle\rangle.
\end{aligned}$$

5 Finally, check that the south-east past yields:

$$\begin{aligned}
&\Big(flat \circ \mathcal{N}[2](mzip) \circ mzip\Big)(\Phi,\Psi)\\
&= \tfrac{2}{3}\big|1|a,0\rangle + 2|a,1\rangle + 1|b,0\rangle + 2|b,1\rangle\big\rangle\\
&\quad + \tfrac{1}{6}\big|2|a,0\rangle + 3|b,1\rangle + 1|a,1\rangle\big\rangle\\
&\quad + \tfrac{1}{6}\big|3|a,1\rangle + 2|b,0\rangle + 1|b,1\rangle\big\rangle.
\end{aligned}$$

This differs from what we get in the first item, via the east-south route.

3.9 Ewens distributions

We conclude this chapter by looking into certain natural multisets and distributions on them that arise in population biology. They were first introduced by Ewens [46]. The aim is to capture situations where new mutations may introduce new elements. One way to model this is to use so-called Hoppe urns, see Example 3.9.7 below. A Hoppe draw is like a Pólya draw — where an additional ball of the same colour as the drawn ball is returned to the urn — but where another ball is added with a new colour, not already occurring in the urn. The latter ball of a new colour corresponds to a new mutation.

We start with the special multisets that are used in this setting.

Definition 3.9.1. An *Ewens* multiset with mean K is a natural multiset $\varphi \in \mathcal{N}(\mathbb{N})$ with $supp(\varphi) \subseteq \{1, 2, \ldots, K\}$ and $mean(\varphi) = K$. For such a multiset φ the mean (average) is defined as:

$$mean(\varphi) \coloneqq \sum_{1 \leq k \leq K} \varphi(k) \cdot k.$$

We shall write $\mathcal{E}(K) \subseteq \mathcal{N}(\mathbb{N})$ for the set of Ewens multisets with mean K. It is easy to see that $\|\varphi\| \leq K$ for $\varphi \in \mathcal{E}(K)$.

For instance, for mean $K = 4$ there are the following Ewens multisets with mean K.

$$4|1\rangle \quad 2|1\rangle + 1|2\rangle \quad 2|2\rangle \quad 1|1\rangle + 1|3\rangle \quad 1|4\rangle. \tag{3.35}$$

Recall that we have been counting lists of coin values in Subsection 1.2.3. There we saw in (1.7) all lists of coins with values $1, 2, 3, 4$ that add up to 4. As we see, the accumulations of these lists are the Ewens multisets of mean 4. In these multiset representations we do not care about the order of the coins.

There are several possible ways of using an Ewens multiset $\varphi = \sum_k n_k|k\rangle \in \mathcal{E}(K)$, so that $\sum_k n_k \cdot k = K$.

- We can think of n_k as the number of coins with value k that are used in φ to form an amount K.
- We can also think of n_k as the number of tables with k customers in an arrangement of K guests in a restaurant, see [6] (or [2, §11.19]) for an early description, and also Exercise 3.9.7 below.
- In genetics, each gene has an 'allelic type' φ, where each n_k is the number of alleles appearing k times.
- An Ewens multiset with mean K is type of a partition of the set $\{1, 2, \ldots, K\}$ in [100]. It tells how many subsets in a partitition have k elements.

In this section we are interested in distributions on Ewens multisets, just like the earlier multinomial / hypergeometric / Pólya draw distributions are distributions on multisets.

We start with a general 'multiplicity count' construction for turning a multiset of size K, on an arbitrary space X, into an Ewens multiset of mean K, with support in $\{1, \ldots, K\}$. This multiplicity count produces a multiset $\sum_k n_k|k\rangle$ where n_k is the number of elements in the original multiset that occur k times. As mentioned above, an Ewens multiset can be seen as a type, capturing how many elements occur how many times. It is a type, as an abstraction, just like the size of a multiset can also be seen as a type. The multiplicity count function calculates this type of an arbitrary multiset, as an Ewens multiset.

Definition 3.9.2. Let X be an arbitrary set. For each number K there is a *multiplicity count* function:

$$\mathcal{N}[K](X) \xrightarrow{mc} \mathcal{E}(K) \qquad \text{given by} \qquad mc(\varphi) \coloneqq \sum_{1 \leq k \leq \|\varphi\|} \left|\varphi^{-1}(k)\right| |k\rangle.$$

Thus, the function mc is defined via the size $|-|$ of inverse image subsets, on $k \geq 1$ as:

$$mc(\varphi)(k) \coloneqq \left|\varphi^{-1}(k)\right| = \left|\{x \in X \mid \varphi(x) = k\}\right|. \tag{3.36}$$

Alternatively, we can use:

$$mc(\varphi) \;=\; \sum_{x \in supp(\varphi)} 1\big|\varphi(x)\big\rangle.$$

We then use sums of multisets implicitly. For instance, for $X = \{a, b, c\}$ and $K = 10$,

$$\begin{aligned} mc\Big(2|a\rangle + 6|b\rangle + 2|c\rangle\Big) \;&=\; 1|2\rangle + 1|6\rangle + 1|2\rangle \;=\; 2|2\rangle + 1|6\rangle \\ mc\Big(2|a\rangle + 3|b\rangle + 5|c\rangle\Big) \;&=\; 1|2\rangle + 1|3\rangle + 1|5\rangle. \end{aligned}$$

We collect some basic results about multiplicity count.

Lemma 3.9.3. *Let a non-empty multiset $\varphi \in \mathcal{N}[K](X)$ be given.*

1 *Taking multiplicity counts is invariant under permutation: when the set X is finite and $\pi \colon X \xrightarrow{\cong} X$ is a permutation (isomorphism), then*

$$mc\Big(\mathcal{D}(\pi)(\varphi)\Big) \;=\; mc(\varphi).$$

2 *For arbitrary $x \in X$,*

$$mc(\varphi + 1|x\rangle) \;=\; mc(\varphi) + 1|\varphi(x)+1\rangle - 1|\varphi(x)\rangle.$$

3 *Similarly, for $x \in supp(\varphi)$,*

$$mc(\varphi - 1|x\rangle) \;=\; \begin{cases} mc(\varphi) - 1|1\rangle & \text{if } \varphi(x) = 1 \\ mc(\varphi) + 1|\varphi(x)-1\rangle - 1|\varphi(x)\rangle & \text{otherwise.} \end{cases}$$

4 $\varphi! \;=\; \displaystyle\prod_{1 \leq k \leq K} (k!)^{mc(\varphi)(k)}$.

5 *For each number K and for each set X with at least K elements, the multiplicity count function $\mathrm{mc} \colon \mathcal{N}[K](X) \to \mathcal{E}(K)$ is surjective.*

Proof. 1 Obvious, since permuting the elements of a multiset does not change the multiplicities that it has.

2 Let $k = \varphi(x)$. If we add another element x to the multiset φ, then the number $mc(\varphi)(k)$ of elements in φ occuring k times is decreased by one, and the number $mc(\varphi)(k + 1)$ of elements occurring $k + 1$ times is increased by one.

3 By a similar argument.

4 By induction on $K \geq 1$. The statement clearly holds when $K = 1$. Next, by

the using the first item,

$$\begin{aligned}
\prod_{1\leq k\leq K+1}(k!)^{mc(\varphi+1|x\rangle)(k)} &= ((K+1)!)^{mc(\varphi+1|x\rangle)(K+1)}\cdot\prod_{1\leq k\leq K}(k!)^{mc(\varphi+1|x\rangle)(k)} \\
&= \begin{cases}(K+1)! & \text{if } \varphi = K|x\rangle \\ \displaystyle\prod_{1\leq k\leq K}(k!)^{mc(\varphi)(k)}\cdot\frac{(\varphi(x)+1)!}{\varphi(x)!} & \text{otherwise}\end{cases} \\
&\overset{\text{(IH)}}{=} \begin{cases}(\varphi+1|x\rangle)\!\!\rceil\!\!\circ & \text{if } \varphi = K|x\rangle \\ \varphi\!\!\rceil\!\!\circ\cdot(\varphi(x)+1) & \text{otherwise}\end{cases} \\
&= (\varphi+1|x\rangle)\!\!\rceil\!\!\circ\,.
\end{aligned}$$

5 Let $\psi = \sum_{1\leq k\leq K} n_k|k\rangle \in \mathcal{E}(K)$ be given. We construct a multiset $\varphi_K \in \mathcal{N}[K](X)$ with $mc(\varphi_K) = \psi$ as follows. We start with the empty multiset φ_0.

For each number k with $1 \leq k \leq K$ we set $\varphi_k = \varphi_{k-1} + k|x_1\rangle + \cdots k|x_{n_k}\rangle$, where $x_1, \ldots, x_{n_k}$ are freshly chosen elements from X, not occurring in φ_{k-1}. This guarantees that φ_k has n_k many elements occurring k times. By construction $mc(\varphi_K) = \psi$. This construction uses $\|\psi\|$ many elements from X. Since $K = \sum_k n_k \cdot k \geq \sum_k n_k = \|\psi\|$, the construction always works for a set X with at least K elements. ❑

In Definition 2.2.4 we have seen draw-and-delete and draw-and-add channels:

$$\mathcal{N}[K](X) \underset{DA}{\overset{DD}{\rightleftarrows}} \mathcal{N}[K+1](X)$$

There are similar channels for Ewens multisets, which we shall write with an extra 'E' for 'Ewens':

$$\mathcal{E}(K) \underset{EDA(t)}{\overset{EDD}{\rightleftarrows}} \mathcal{E}(K+1) \tag{3.37}$$

Notice that the draw-add map $EDA(t)$ involves a parameter t. It is a positive real number, which is used as mutation rate in population genetics. Its role will become clear below.

These maps in (3.37) are channels. Thus, $EDA(t)$ and EDD produce a distribution over Ewens multisets, with increased or decreased mean. These increases and descreses are a bit subtle, because we have to make sure that the right mean emerges. The trick is to shift one occurrence from $|i\rangle$ to $|i+1\rangle$, or

to shift in the other direction. This is achieved in the formulations in the next definition. The deletion construction comes from [100, 101] and the addition construction is used in the 'Chinese restaurant' illustration, see Exercise 3.9.7 below. Both constructions can be found in [27, §3.1 and §4.5].

Later on, in Proposition 6.6.6, we shall show that the channels *EDD* and $EDA(t)$ are closely related, in the sense that they are each other's 'daggers'.

Definition 3.9.4. For Ewens multisets $\varphi \in \mathcal{E}(K)$ and $\psi \in \mathcal{E}(K{+}1)$ and for $t > 0$ define:

$$\begin{aligned} EDA(t)(\varphi) &\coloneqq \frac{t}{K+t}\big|\varphi + 1|1\rangle\big\rangle + \sum_{1\leq k\leq K} \frac{\varphi(k)\cdot k}{K+t}\big|\varphi - 1|k\rangle + 1|k{+}1\rangle\big\rangle \\ EDD(\psi) &\coloneqq \frac{\psi(1)}{K+1}\big|\psi - 1|1\rangle\big\rangle + \sum_{2\leq k\leq K+1} \frac{\psi(k)\cdot k}{K+1}\big|\psi + 1|k{-}1\rangle - 1|k\rangle\big\rangle \end{aligned}$$

For instance, for $\varphi = 1|1\rangle + 2|3\rangle + 1|6\rangle \in \mathcal{E}(13)$ we get:

$$\begin{aligned} EDA(1)(\varphi) &= \tfrac{1}{14}\big|2|1\rangle + 2|3\rangle + 1|6\rangle\big\rangle + \tfrac{1}{14}\big|1|2\rangle + 2|3\rangle + 1|6\rangle\big\rangle \\ &\qquad + \tfrac{3}{7}\big|1|1\rangle + 1|3\rangle + 1|4\rangle + 1|6\rangle\big\rangle + \tfrac{3}{7}\big|1|1\rangle + 2|3\rangle + 1|7\rangle\big\rangle \\ EDD(\varphi) &= \tfrac{6}{13}\big|2|3\rangle + 1|6\rangle\big\rangle + \tfrac{6}{13}\big|1|1\rangle + 1|2\rangle + 1|3\rangle + 1|6\rangle\big\rangle \\ &\qquad + \tfrac{1}{13}\big|1|1\rangle + 2|3\rangle + 1|5\rangle\big\rangle. \end{aligned}$$

The multiplicity count function *mc* from Definition 3.9.2 interacts smoothly with draw-delete. This is what we show first.

Lemma 3.9.5. *Multiplicity count commutes with draw-delete, as in:*

$$\begin{array}{ccc} \mathcal{N}[K](X) & \xleftarrow{DD} & \mathcal{N}[K+1](X) \\ {\scriptstyle mc}\downarrow & & \downarrow{\scriptstyle mc} \\ \mathcal{E}(K) & \xleftarrow[EDD]{} & \mathcal{E}(K+1) \end{array}$$

In this diagram the function mc *is used as deterministic channel.*

Proof. We recall from Definition 2.2.4 that for $\psi \in \mathcal{N}[K+1](X)$ one has:

$$DD(\psi) \coloneqq \sum_{x\in supp(\psi)} \frac{\psi(x)}{K+1}\,\big|\psi - 1|x\rangle\big\rangle$$

Then, via Lemma 3.9.3 (3),

$$\begin{aligned}
&\mathcal{D}(mc)\Big(DD(\psi)\Big) \\
&= \sum_{x\in supp(\psi)} \frac{\psi(x)}{K+1}\,\big|mc(\psi-1|x\rangle)\big\rangle \\
&= \sum_{x\in supp(\psi)} \frac{\psi(x)}{K+1}\begin{cases} \big|mc(\psi)-1|1\,\rangle\big\rangle & \text{if } \psi(x)=1 \\ \big|mc(\psi)+1|\psi(x)-1\,\rangle-1|\psi(x)\rangle\big\rangle & \text{otherwise.}\end{cases} \\
&= \sum_{x,\,\psi(x)=1} \frac{\psi(x)}{K+1}\big|mc(\psi)-1|1\,\rangle\big\rangle + \sum_{x,\,\psi(x)>1} \frac{\psi(x)}{K+1}\big|mc(\psi)+1|\psi(x)-1\,\rangle-1|\psi(x)\rangle\big\rangle \\
&= \frac{mc(\psi)(1)}{K+1}\big|mc(\psi)-1|1\,\rangle\big\rangle + \sum_{2\le k\le K+1} \frac{\sum_{x,\,\psi(x)=k}\psi(x)}{K+1}\big|mc(\psi)+1|k-1\,\rangle-1|k\,\rangle\big\rangle \\
&= \frac{mc(\psi)(1)}{K+1}\big|mc(\psi)-1|1\,\rangle\big\rangle + \sum_{2\le k\le K+1} \frac{mc(\psi)(k)\cdot k}{K+1}\big|mc(\psi)+1|k-1\,\rangle+1|k\,\rangle\big\rangle \\
&= EDD\Big(mc(\psi)\Big).
\end{aligned}$$

This completes the proof. ❑

We now come to a crucial channel $ew[K]\colon \mathbb{R}_{>0} \nrightarrow \mathcal{E}(K)$ called after Ewens who introduced the distributions $ew[K](t) \in \mathcal{D}(\mathcal{E}(K))$ in [46]. These distributions will be introduced by induction on K. Subsequently, an explicit formula is obtained, for the special case when the parameter t is a natural number.

Definition 3.9.6. Fix a parameter $t \in \mathbb{R}_{>0}$ and define by induction:

$$\begin{aligned} ew[1](t) &:= 1\big|1|1\,\rangle\big\rangle \\ ew[K+1](t) &:= EDA(t) \gg\!= ew[K](t). \end{aligned} \tag{3.38}$$

The first distribution $ew[1](t) = 1\big|1|1\,\rangle\big\rangle \in \mathcal{D}(\mathcal{E}(1))$ contains a lot of 1's. It is the singleton distribution for the singleton Ewens multiset $1|1\,\rangle$ on $\{1\}$ with mean 1. In this first step the parameter $t \in \mathbb{R}_{>0}$ does not play a role. The two next steps yield:

$$\begin{aligned} ew[2](t) \;=\; EDA(t) \gg\!= 1\big|1|1\,\rangle\big\rangle \;&=\; EDA(t)(1|1\,\rangle) \\ &= \frac{t}{1+t}\big|2|1\,\rangle\big\rangle + \frac{1}{1+t}\big|1|2\,\rangle\big\rangle. \end{aligned}$$

And:

$$\begin{aligned}
&ew[3](t)\\
&= EDA(t) \gg\!= \left(\frac{t}{1+t}\big|2|1\rangle\big\rangle + \frac{1}{1+t}\big|1|2\rangle\big\rangle\right)\\
&= \frac{t}{1+t} \cdot EDA(t)(2|1\rangle) + \frac{1}{1+t} \cdot EDA(t)(1|2\rangle)\\
&= \frac{t}{1+t} \cdot \frac{t}{2+t}\big|3|1\rangle\big\rangle + \frac{t}{1+t} \cdot \frac{2}{2+t}\big|1|1\rangle + 1|2\rangle\big\rangle\\
&\qquad + \frac{1}{1+t} \cdot \frac{t}{2+t}\big|1|1\rangle + 1|2\rangle\big\rangle + \frac{1}{1+t} \cdot \frac{2}{2+t}\big|1|3\rangle\big\rangle\\
&= \frac{t^2}{(1+t)(2+t)}\big|3|1\rangle\big\rangle + \frac{3t}{(1+t)(2+t)}\big|1|1\rangle + 1|2\rangle\big\rangle + \frac{2}{(1+t)(2+t)}\big|1|3\rangle\big\rangle.
\end{aligned}$$

There are several alternative ways to describe Ewens distributions, certainly when the mutation rate t is a natural number. The first alternative involves drawing from an urn; it justifies putting the topic of Ewens multisets / distributions in this chapter on drawing. The following drawing method was introduced by Hoppe [65]. It extends Pólya urns with the option of introducing a ball with a new colour. It corresponds to the idea of a new mutation in genetics.

Example 3.9.7. For numbers $t \in \mathbb{R}_{>0}$ and $K \in \mathbb{N}_{>0}$ define the *Hoppe* channel:

$$\mathcal{N}[K](\{1,\ldots,K\}) \xrightarrow{hop[K](t)} \mathcal{N}[K+1](\{1,\ldots,K+1\})$$

on $\varphi \in \mathcal{N}[K](\{1,\ldots,K\})$ as:

$$hop[K](t)(\varphi) \coloneqq \sum_{1\leq k\leq K} \frac{\varphi(k)}{K+t}\big|\varphi + 1|k\rangle\big\rangle + \frac{t}{K+t}\big|\varphi + 1|K+1\rangle\big\rangle. \tag{3.39}$$

On the right-hand side we implicitly promote φ to a multiset on $\{1,\ldots,K+1\}$. In these Hoppe draws we use φ as an urn filled with coloured balls, where the colours are numbers $1,\ldots,K$. Drawn elements of colour k are returned, together with an extra copy of the same colour, like in Pólya urns. But what's different is that an additional ball of a new colour $K+1$ is added, in the term $\varphi + 1|K+1\rangle$ in (3.39).

These Hoppe draws can be repeated and give, starting from $1|1\rangle \in \mathcal{M}[1](\{1\}$

the following distributions. For convenience, we omit the parameter K.

$$\begin{aligned} h(t)(1|1\rangle) &= \frac{1}{1+t}\big|2|1\rangle\big\rangle + \frac{t}{1+t}\big|1|1\rangle + 1|2\rangle\big\rangle \\ (h(t) \circ h(t))(1|1\rangle) &= \frac{2}{(1+t)(2+t)}\big|3|1\rangle\big\rangle + \frac{t}{(1+t)(2+t)}\big|2|1\rangle + 1|3\rangle\big\rangle \\ &\quad + \frac{t}{(1+t)(2+t)}\big|2|1\rangle + 1|2\rangle\big\rangle + \frac{t}{(1+t)(2+t)}\big|1|1\rangle + 2|2\rangle\big\rangle \\ &\quad + \frac{t^2}{(1+t)(2+t)}\big|1|1\rangle + 1|2\rangle + 1|3\rangle\big\rangle. \end{aligned}$$

Notice that the multisets in these distributions are not Ewens multisets. But we can turn them into such Ewens multiset via multiplicity count *mc* from Definition 3.9.2. Interestingly, we then get the Ewens distributions from Definition 3.9.6.

The following result can be seen as an analogue of Lemma 3.9.5, for addition instead of for deletion. It is however more restricted, since it involves sample spaces of the form $\{1, \ldots, K\}$, and not arbitrary sets.

Lemma 3.9.8. *Multiplicity count* mc *connects Hoppe and add-delete on Ewens multisets, as in:*

$$\begin{array}{ccc} \mathcal{N}[K](\{1,\ldots,K\}) & \xrightarrow{hop(t)} & \mathcal{N}[K+1](\{1,\ldots,K+1\}) \\ {\scriptstyle mc}\downarrow & & \downarrow{\scriptstyle mc} \\ \mathcal{E}(K) & \xrightarrow[EDA(t)]{} & \mathcal{E}(K+1) \end{array}$$

This rectangle commutes for arbitrary $t \in \mathbb{R}_{>0}$.

Proof. Let a multiset $\varphi \in \mathcal{N}[K](\{1, \ldots, K\})$ be given. Via Lemma 3.9.3 (2):

$$\begin{aligned} &\mathcal{D}(mc)\Big(hop(t)(\varphi)\Big) \\ &\overset{(3.39)}{=} \sum_{1\leq \ell \leq K} \frac{\varphi(\ell)}{K+t}\big|mc(\varphi + 1|\ell\rangle)\big\rangle + \frac{t}{K+t}\big|mc(\varphi + 1|K+1\rangle)\big\rangle \\ &= \sum_{1\leq \ell \leq K} \frac{\varphi(\ell)}{K+t}\big|mc(\varphi) + 1|\varphi(\ell)+1\rangle - 1|\varphi(\ell)\rangle\big\rangle + \frac{t}{K+t}\big|mc(\varphi) + 1|1\rangle\big\rangle \\ &\overset{(*)}{=} \frac{t}{K+t}\big|mc(\varphi) + 1|1\rangle\big\rangle + \sum_{1\leq k \leq K} \frac{mc(\varphi)(k)\cdot k}{K+t}\big|mc(\varphi) - 1|k\rangle + 1|k+1\rangle\big\rangle \\ &= EDA(t)\Big(mc(\varphi)\Big). \end{aligned}$$

The marked equation $\overset{(*)}{=}$ holds since by definition, $mc(\varphi)(k) = \big|\{\ell \mid \varphi(\ell) = k\}\big|$, see (3.36). ❑

Corollary 3.9.9. *Ewens distributions can also be obtained via repeated Hoppe draws followed by multiplicity count: for $t \in \mathbb{R}_{>0}$ and $K \in \mathbb{N}_{>0}$, one has:*

$$ew[K](t) = \Big(mc \circ \underbrace{hop(t) \circ \cdots \circ hop(t)}_{K-1 \text{ times}}\Big)(1|1\rangle).$$

Proof. By induction on $K \geq 1$, using that $mc(1|1\rangle) = 1|1\rangle$ and Lemma 3.9.8. ❑

It turns out that there is a single formula for Ewens distributions when the mutation rate t is a natural number. This formula is often called *Ewens sampling formula*.

Theorem 3.9.10. *When the parameter $t > 0$ is a natural, the Ewens distribution $ew[K](t) \in \mathcal{D}(\mathcal{E}(K))$ can be described by:*

$$ew[K](t) = \sum_{\varphi \in \mathcal{E}(K)} \frac{\prod_{1 \leq k \leq K} \left(\frac{t}{k}\right)^{\varphi(k)}}{\varphi! \cdot \binom{t}{K}} \big|\varphi\rangle. \tag{3.40}$$

Proof. We fix $t \in \mathbb{N}_{>0}$ and use induction on $K \geq 1$. When $K = 1$ the only possible Ewens multiset is $\varphi = 1|1\rangle \in \mathcal{E}(1)$, so that:

$$\frac{\left(\frac{t}{1}\right)^1}{1! \cdot \binom{t}{1}} \big|1|1\rangle\big\rangle = \frac{t}{\binom{t}{1}} \big|1|1\rangle\big\rangle = \frac{t}{t} \big|1|1\rangle\big\rangle = 1\big|1|1\rangle\big\rangle \overset{(3.38)}{=} ew[1](t).$$

For the induction step we assume that Equation (3.40) holds for all numbers below K. We first note that we can write for $\varphi \in \mathcal{E}(K)$ and $\psi \in \mathcal{E}(K+1)$,

$$\begin{aligned} EDA(t)(\varphi)(\psi) &= \begin{cases} \dfrac{t}{K+t} & \text{if } \psi = \varphi + 1|1\rangle \\ \dfrac{\varphi(k) \cdot k}{K+t} & \text{if } \psi = \varphi - 1|k\rangle + 1|k+1\rangle \end{cases} \\ &= \begin{cases} \dfrac{t}{K+t} & \text{if } \varphi = \psi - 1|1\rangle \\ \dfrac{(\psi(k)+1) \cdot k}{K+t} & \text{if } \varphi = \psi + 1|k\rangle - 1|k+1\rangle \end{cases} \end{aligned}$$

where $1 \leq k \leq K$. Hence:

$$\begin{aligned}
ew[K+1](t)(\psi) &\overset{(3.38)}{=} \Big(EDA(t) \gg\!= ew[K](t)\Big)(\psi) \\
&= \sum_{\varphi \in \mathcal{E}(K)} EDA(t)(\varphi)(\psi) \cdot ew[K](t)(\varphi) \\
&= \frac{t}{K+t} \cdot ew[K](t)(\psi - 1|1\rangle) \qquad (\text{if } \psi(1) > 0) \\
&\quad + \sum_{1 \leq k \leq K} \frac{(\psi(k)+1) \cdot k}{K+t} \cdot ew[K](t)(\psi + 1|k\rangle - 1|k+1\rangle) \\
&\overset{(\text{IH})}{=} \frac{t}{K+t} \cdot \frac{\prod_{1 \leq \ell \leq K} \left(\frac{t}{\ell}\right)^{(\psi - 1|1\rangle)(\ell)}}{(\psi - 1|1\rangle)! \cdot \binom{t}{K}} \\
&\quad + \sum_{1 \leq k \leq K} \frac{(\psi(k)+1) \cdot k}{K+t} \cdot \frac{\prod_{1 \leq \ell \leq K} \left(\frac{t}{\ell}\right)^{(\psi + 1|k\rangle - 1|k+1\rangle)(\ell)}}{(\psi + 1|k\rangle - 1|k+1\rangle)! \cdot \binom{t}{K}} \\
&= \frac{\psi(1)}{K+1} \cdot \frac{\prod_{1 \leq \ell \leq K+1} \left(\frac{t}{\ell}\right)^{\psi(\ell)}}{\psi! \cdot \binom{t}{K+1}} \\
&\quad + \sum_{1 \leq k \leq K} \frac{\psi(k+1) \cdot (k+1)}{K+1} \cdot \frac{\prod_{1 \leq \ell \leq K+1} \left(\frac{t}{\ell}\right)^{\psi(\ell)}}{\psi! \cdot \binom{t}{K+1}} \\
&= \frac{\psi(1) + \sum_{1 \leq k \leq K} \psi(k+1) \cdot (k+1)}{K+1} \cdot \frac{\prod_{1 \leq \ell \leq K+1} \left(\frac{t}{\ell}\right)^{\psi(\ell)}}{\psi! \cdot \binom{t}{K+1}} \\
&= \frac{\prod_{1 \leq \ell \leq K+1} \left(\frac{t}{\ell}\right)^{\psi(\ell)}}{\psi! \cdot \binom{t}{K+1}}.
\end{aligned}$$

❑

This result has a number of consequences. We start with an obvious one, reformulating that the probabilities in (3.40) add up to one. This result resembles earlier results, in Lemma 1.6.2 and Proposition 1.7.3, underlying the multinomial and Pólya distributions.

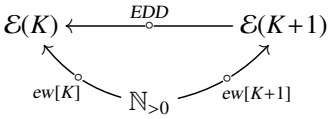
Corollary 3.9.11. *For $K, t \in \mathbb{N}_{>0}$,*

$$\sum_{\varphi \in \mathcal{E}(K)} \frac{t^{\|\varphi\|}}{\varphi! \cdot \prod_{1 \leq k \leq K} k^{\varphi(k)}} = \binom{t}{K}.$$

❑

By construction, in Definition 3.9.6, Ewens channels commute with Ewens draw-addition maps *EDA*. With their explicit formulation (3.40) we can show that Ewens channels, when restricted to positive natural numbers, commute with Ewens draw-delete maps *EDD*. Thus these channels form a cone for the infinite chain of draw-delete maps. This resembles the situation for multinomial channels in Proposition 3.3.8.

Corollary 3.9.12. *For $K \geq 1$ the following triangles commute:*



Recall from Exercise 2.2.10 that there are also 'fixed point' distributions for the ordinary (non-Ewens) draw-delete/add maps, but the ones given there are not as rich as the Ewens distributions.

Proof. We first extract the following characterisation from the formulation of *EDD* in Definition 3.9.4: for $\psi \in \mathcal{E}(K+1)$ and $\varphi \in \mathcal{E}(K)$,

$$\begin{aligned} EDD(\psi)(\varphi) &= \begin{cases} \dfrac{\psi(1)}{K+1} & \text{if } \varphi = \psi - 1|1\rangle \\ \dfrac{\psi(k)\cdot k}{K+1} & \text{if } \varphi = \psi + 1|k-1\rangle - 1|k\rangle \end{cases} \\ &= \begin{cases} \dfrac{\varphi(1)+1}{K+1} & \text{if } \psi = \varphi + 1|1\rangle \\ \dfrac{(\varphi(k)+1)\cdot k}{K+1} & \text{if } \psi = \varphi - 1|k-1\rangle + 1|k\rangle \end{cases} \end{aligned}$$

where $2 \leq k \leq K+1$. Hence:

$$\begin{aligned} &\Big(EDD \gg\!= ew[K](t)\Big)(\varphi) \\ &= \sum_{\psi \in \mathcal{E}(K+1)} EDD(\psi)(\varphi) \cdot ew[K+1](t)(\psi) \\ &= \frac{\varphi(1)+1}{K+1} \cdot ew[K+1](t)(\varphi + 1|1\rangle) \\ &\qquad + \sum_{1\leq k\leq K} \frac{(\varphi(k+1)+1)\cdot(k+1)}{K+1} \cdot ew[K+1](t)(\varphi - 1|k\rangle + 1|k+1\rangle) \\ &\overset{(3.40)}{=} \frac{\varphi(1)+1}{K+1} \cdot \frac{\prod_{1\leq \ell\leq K+1} \binom{t}{\ell}^{(\varphi+1|1\rangle)(\ell)}}{(\varphi+1|1\rangle)!_\circ \cdot \binom{t}{K+1}} \\ &\qquad + \sum_{1\leq k\leq K} \frac{(\varphi(k+1)+1)\cdot(k+1)}{K+1} \cdot \frac{\prod_{1\leq \ell\leq K+1} \binom{t}{\ell}^{(\varphi-1|k\rangle+1|k+1\rangle)(\ell)}}{(\varphi-1|k\rangle+1|k+1\rangle)!_\circ \cdot \binom{t}{K+1}} \\ &= \frac{t}{K+t} \cdot \frac{\prod_{1\leq \ell\leq K} \binom{t}{\ell}^{\varphi(\ell)}}{\varphi!_\circ \cdot \binom{t}{K}} + \sum_{1\leq k\leq K} \frac{\varphi(k)\cdot k}{K+t} \cdot \frac{\prod_{1\leq \ell\leq K} \binom{t}{\ell}^{\varphi(\ell)}}{\varphi!_\circ \cdot \binom{t}{K}} \\ &= \frac{t}{K+t} \cdot ew[K](t)(\varphi) + \frac{\sum_{1\leq k\leq K} \varphi(k)\cdot k}{K+t} \cdot ew[K](t)(\varphi) \\ &= ew[K](t)(\varphi). \end{aligned}$$

❑

We consider another consequence of Theorem 3.9.10. It fills a gap that

we left open at the very beginning of this book, namely the proof of Theorem 1.2.8. There we looked at sums $sum(\ell)$ and products $prod(\ell)$ of sequences $\ell \in \mathcal{L}(\mathbb{N}_{>0})$ of positive natural numbers. By accumulating such lists into multisets we get a connection with Ewens multisets, namely:

$$sum(\ell) = K \iff acc(\ell) \in \mathcal{E}(K). \tag{3.41}$$

Thus if we arrange the multisets (with mean K) in the Ewens distribution we get a distribution on lists that sum to a certain number K.

Corollary 3.9.13. *For $t, K \in \mathbb{N}_{>0}$,*

$$arr \gg\!= ew[K](t) = \sum_{\ell \in sum^{-1}(K)} \frac{t^{\|\varphi\|}}{\|\ell\|! \cdot prod(\ell) \cdot \binom{t}{K}} \big|\ell\big\rangle.$$

In particular, for $t = 1$ we get a distribution:

$$arr \gg\!= ew[K](1) = \sum_{\ell \in sum^{-1}(K)} \frac{1}{\|\ell\|! \cdot prod(\ell)} \big|\ell\big\rangle,$$

in which all probabilities add up to one. This is what was claimed earlier in Theorem 1.2.8, at the time without proof.

Proof. We elaborate:

$$\begin{aligned}
arr \gg\!= ew[K](t) &= \sum_{\ell \in \mathcal{L}(\{1,\ldots,K\})} \left(\sum_{\varphi \in \mathcal{E}(K)} arr(\varphi)(\ell) \cdot ew[K](t)(\varphi) \right) \big|\ell\big\rangle \\
&\overset{(3.40)}{=} \sum_{\ell \in \mathcal{L}(\{1,\ldots,K\})} \left(\sum_{\varphi \in \mathcal{E}(K),\, acc(\ell) = \varphi} \frac{1}{(\varphi)} \cdot \frac{\prod_{1 \leq k \leq K} \binom{t}{k}^{\varphi(k)}}{\varphi! \cdot \binom{t}{K}} \right) \big|\ell\big\rangle \\
&\overset{(*)}{=} \sum_{\ell \in \mathcal{L}(\{1,\ldots,K\})} \left(\sum_{\varphi \in \mathcal{E}(K),\, acc(\ell) = \varphi} \frac{1}{\|\varphi\|!} \cdot \frac{t^{\|\varphi\|}}{prod(\ell) \cdot \binom{t}{K}} \right) \big|\ell\big\rangle \\
&\overset{(3.41)}{=} \sum_{\ell \in sum^{-1}(K)} \frac{t^{\|\varphi\|}}{\|\ell\|! \cdot prod(\ell) \cdot \binom{t}{K}} \big|\ell\big\rangle.
\end{aligned}$$

The marked equation holds because if $acc(\ell) = \varphi \in \mathcal{E}(K)$, then:

$$\frac{1}{prod(\ell)} = \prod_{1 \leq k \leq K} \Big(\frac{1}{k}\Big)^{\varphi(k)}.$$

For instance, for $\ell = [1, 2, 2, 3]$ with $\varphi = acc(\ell) = 1|1\rangle + 2|2\rangle + 1|3\rangle \in \mathcal{E}(8)$ we have:

$$\frac{1}{prod(\ell)} = \frac{1}{1 \cdot 2 \cdot 2 \cdot 3} = \frac{1}{12} = \frac{1}{1} \cdot \Big(\frac{1}{2}\Big)^2 \cdot \frac{1}{3} = \prod_{1 \leq k \leq K} \Big(\frac{1}{k}\Big)^{\varphi(k)}. \qquad \square$$

Exercises

3.9.1 Check that the sets of Ewens multisets $\mathcal{E}(1)$, $\mathcal{E}(2)$, $\mathcal{E}(3)$, $\mathcal{E}(5)$ have, respectively, 1, 2, 3, 5 and 7 elements; list them all.

3.9.2 Use the construction from the proof of Lemma 3.9.3 (5) to construct for $\varphi = 2|1\rangle + 1|4\rangle + 2|5\rangle \in \mathcal{E}(16)$ a multiset ψ with $mc(\psi) = \varphi$.

3.9.3 Use the isomorphism $\mathbb{N} \cong \mathcal{N}(1)$ to describe the mean function from Definition 3.9.1 as an instance of flattening.

3.9.4 Show that for $\ell \in \mathcal{L}(\mathbb{N}_{>0})$ one has:

$$mean(acc(\ell)) \;=\; sum(\ell).$$

3.9.5 Recall from Lemma 1.2.7 that for $N \in \mathbb{N}_{>0}$ there are 2^{N-1} many lists $\ell \in \mathcal{L}(\mathbb{N}_{>0})$ with $sum(\ell) = N$. Hence we have a uniform distribution:

$$unif_{sum^{-1}(N)} \;=\; \sum_{\ell \in sum^{-1}(N)} \frac{1}{2^{N-1}} \big|\ell\big\rangle.$$

Via accumulation we get a distribution in $\mathcal{E}(N)$, namely:

$$\mathcal{D}(acc)\Big(unif_{sum^{-1}(N)}\Big) \;=\; \sum_{\ell \in sum^{-1}(N)} \frac{1}{2^{N-1}} \big|acc(\ell)\big\rangle.$$

Use (1.7) to show that for $N = 4$ this construction yields

$$\begin{aligned}&\mathcal{D}(acc)\Big(unif_{sum^{-1}(4)}\Big)\\ &= \tfrac{1}{8}\big|4|1\rangle\big\rangle + \tfrac{3}{8}\big|2|1\rangle + 1|2\rangle\big\rangle + \tfrac{1}{8}\big|2|2\rangle\big\rangle + \tfrac{1}{4}\big|1|1\rangle + 1|3\rangle\big\rangle + \tfrac{1}{8}\big|1|4\rangle\big\rangle.\end{aligned}$$

(This is not *ew*[4](1).)

3.9.6 Take $\varphi = 3|a\rangle + 1|b\rangle + 1|c\rangle$. Show that:

$$\begin{aligned}&EDA(1)\Big(mc(\varphi)\Big)\\ &= \tfrac{1}{6}\big|3|1\rangle + 1|3\rangle\big\rangle + \tfrac{1}{3}\big|1|1\rangle + 1|2\rangle + 1|3\rangle\big\rangle + \tfrac{1}{2}\big|2|1\rangle + 1|4\rangle\big\rangle\end{aligned}$$

and:

$$\mathcal{D}(mc)\Big(DA(\varphi)\Big) \;=\; \tfrac{2}{5}\big|1|1\rangle + 1|2\rangle + 1|3\rangle\big\rangle + \tfrac{3}{5}\big|2|1\rangle + 1|4\rangle\big\rangle.$$

This demonstrates that a "draw" analogue of Lemma 3.9.5 does not work, at least not for mutatation rate $t = 1$.

3.9.7 The following Chinese Restaurant process is a standard construction in the theory of Ewens distributions, see *e.g.* [6] or [27, §4.5]. We quote from the latter.

> Imagine a restaurant in which customers are labeled according to the order in which they arrive: the first customer is labeled 1, the second is labeled 2, and so on. If the first K customers are seated at $m \geq 1$ different tables, the $(K + 1)$st customer sits

- at a table occupied by $k \geq 1$ customers with probability $k/(t + K)$ or
- at an unoccupied table with probability $t/(t + K)$,

where $t > 0$.

(We have adapted the symbols to make them consistent with the notation used here, *e.g.* with t instead of θ.)

Check that the draw-add channel $EDA(t)$ from Definition 3.9.4 captures the above process. Keep in mind that an Ewens multiset $\varphi = \sum_k n_k|k\rangle \in \mathcal{E}(K)$ describes a seating arrangement with n_k tables have k customers.

3.9.8 Give a proof of the equivalence in (3.41).

3.9.9 Show that for $\varphi \in \mathcal{E}(K)$,

$$\sum_{i \in supp(\varphi)} ew[K](1)(\varphi - 1|i\rangle) \;=\; K \cdot ew[K](1)(\varphi).$$

This property is called non-interference, see [100, 27]. It resembles the recurrence relation for multinomials in Exercise 3.3.6.

3.9.10 Recall the infinite multinomial $\overset{\infty}{mn}[K] \colon \mathcal{D}_\infty(\mathbb{N}_{>0}) \to \mathcal{D}_\infty(\mathcal{N}[K](\mathbb{N}_{>0}))$ from Exercise 3.3.16. We define:

$$\varepsilon[K] \;\coloneqq\; \Big(\mathcal{D}_\infty(\mathbb{N}_{>0}) \xrightarrow{\overset{\infty}{mn}[K]} \mathcal{D}_\infty(\mathcal{N}[K](\mathbb{N}_{>0})) \xrightarrow{\mathcal{D}(mc)} \mathcal{D}_\infty(\mathcal{E}(K))\Big)$$

1 Check that for $\rho \in \mathcal{D}_\infty(\mathbb{N}_{>0})$ one actually gets $\varepsilon[K](\rho)$ in $\mathcal{D}(\mathcal{E}(K))$ instead of in $\mathcal{D}_\infty(\mathcal{E}(K))$.

2 Show that $EDD \circ \varepsilon[K+1] = \varepsilon[K]$, so that the ε's form a cone for the infinite chain of EDD maps.

Such a cone is called a partition structure in [100, 101]. The above $\varepsilon[K]$ maps correspond to the maps ϕ_K in [100, (2.10)]; to see this Lemma 3.9.3 (4) is useful. They carry a De Finetti structure via the Poisson-Dirichlet distribution.

4

Observables and validity

We have seen (discrete probability) distributions as formal convex sums $\sum_i r_i|x_i\rangle$ in $\mathcal{D}(X)$ and (probabilistic) channels $X \nrightarrow Y$, as functions $X \to \mathcal{D}(Y)$, describing probabilistic states and computations. This section develops the tools to reason about such distributions and channels, via what are called *observables*. They are functions from a set / sample space X to (a subset of) the real numbers $\mathbb{R}$ that associate some 'observable' numerical information with an element $x \in X$. The following table gives an overview of terminology and types, where X is a set (used as sample space).

name	**type**
observable / utility function	$X \to \mathbb{R}$
factor / potential function	$X \to \mathbb{R}_{\geq 0}$
(fuzzy) predicate / (soft / uncertain) evidence	$X \to [0, 1]$
sharp predicate / event	$X \to \{0, 1\}$

(4.1)

We shall use the term of 'observable' as generic expression for all the entries in this table. A function $X \to \mathbb{R}$ is thus the most general type of observable, and a sharp predicate $X \to \{0, 1\}$ is the most specific one. Predicates are the most appropriate observable for probabilistic logical reasoning. Often attention is restricted to subsets $U \subseteq X$ as predicates (or events [144]), but here, in this book, the fuzzy versions $X \to [0, 1]$ are the default. Such fuzzy predicates may also be called belief functions — or effects, in a quantum setting. A technical reason using fuzzy, $[0, 1]$-valued predicates is that these fuzzy predicates are closed under predicate transformation $=\!\!\ll$, and the sharp predicates are not, see below for details.

The commonly used notion of *random variable* can now be described as a pair, consisting of an observable together with a state (distribution), both on the same sample space. Often, this state is left implicit; it may be obvious in a particular context what it is. But it may sometimes also be confusing, for instance when we deal with two random variables and we need to make explicit wether they involve different states, or share a state. Like elsewhere in this book, we like to be explicit about the state(s) that we are using.

This chapter starts with the definition of what can be seen as probabilistic truth $\omega \models p$, namely the validity of an observable p in a state ω. It is the expected value of p in ω. We shall see that many basic concepts can be defined in terms of validity, including mean, average, entropy, distance, (co)variance. The algebraic, logical and categorical structure of the various observables in Table 4.1 will be investigated in Section 4.2.

In the previous chapter we have seen that a state ω on the domain X of a channel $c\colon X \multimap Y$ can be transformed into a state $c \gg\!= \omega$ on the channel's codomain Y. Analogous to such state transformation $\gg\!=$ there is also observable transformation $=\!\ll$, acting in the opposite direction: for an observable q on the codomain Y of a channel $c\colon X \multimap Y$, there is a transformed observable $c =\!\ll q$ on the domain X. When $=\!\ll$ is applied to predicates, it is called predicate transformation. It is a basic operation in programming logics that will be investigated in this section. These transformations in different directions are an aspect of the duality between states and predicates. These two transformations correspond to Schrödinger's (forward) Heisenberg's (backward) approach in quantum foundations, see [74]. In the end, in Section 4.5, validity is used to give (dual) formulations of distances between states and between predicates. Via this distance function we can make important properties precise, like: each distribution can be approximated, with arbitrary precision, via frequentist learning of natural multisets. Technically, the 'rational' distributions of the form $\mathit{Flrn}(\varphi)$, for natural multisets φ, form a dense subset of the (metric) space of all distributions.

4.1 Validity

This section introduces the basic facts and terminology for observables, as described in Table 4.1 and defines their validity in a state. Recall that we write Y^X for the set of functions from X to Y. We will use notations:

- $\mathit{Obs}(X) \coloneqq \mathbb{R}^X$ for the set of observables on a set X;
- $\mathit{Fact}(X) \coloneqq (\mathbb{R}_{\geq 0})^X$ for the factors on X;

- $Pred(X) \coloneqq [0,1]^X$ for the predicates on X;
- $SPred(X) \coloneqq \{0,1\}^X$ for the sharp predicates (events) on X.

There are inclusions:

$$SPred(X) \subseteq Pred(X) \subseteq Fact(X) \subseteq Obs(X).$$

The first set $SPred(X) = \{0,1\}^X$ of sharp predicates can be identified with the powerset $\mathcal{P}(X)$ of subsets of X, see below. We first define some special observables.

Definition 4.1.1. Let X be an arbitrary set.

1 For a subset $U \subseteq X$ we write $\mathbf{1}_U \colon X \to \{0,1\}$ for the *characteristic function* of U, defined as:

$$\mathbf{1}_U(x) \coloneqq \begin{cases} 1 & \text{for } x \in U \\ 0 & \text{otherwise.} \end{cases}$$

This function $\mathbf{1}_U \colon X \to [0,1]$ is the (sharp) predicate associated with the subset $U \subseteq X$.

2 We use special notation for two extreme cases $U = X$ and $U = \emptyset$, giving the *truth* predicate $\mathbf{1} \colon X \to [0,1]$ and the *falsity* predicate $\mathbf{0} \colon X \to [0,1]$ on X. Explicitly:

$$\mathbf{1} \coloneqq \mathbf{1}_X \quad \text{and} \quad \mathbf{0} \coloneqq \mathbf{1}_\emptyset \qquad \text{so that} \qquad \mathbf{1}(x) = 1 \quad \text{and} \quad \mathbf{0}(x) = 0,$$

for all $x \in X$.

3 For a singleton subset $\{x\}$ we simply write $\mathbf{1}_x$ for $\mathbf{1}_{\{x\}}$. Such functions $\mathbf{1}_x \colon X \to [0,1]$ are also called *point predicates*, where the element $x \in X$ is seen as a point.

4 There is a (sharp) equality predicate $Eq \colon X \times X \to [0,1]$ defined in the obvious way as:

$$Eq(x,x') \coloneqq \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise.} \end{cases}$$

5 If the set X can be mapped into $\mathbb{R}$ in an obvious manner, we write this inclusion function as $incl_X \colon X \hookrightarrow \mathbb{R}$ and may consider it as an observable on X. This applies for instance if $X = n = \{0, 1, \ldots, n-1\}$.

This mapping $U \mapsto \mathbf{1}_U$ forms the isomorphism $\mathcal{P}(X) \xrightarrow{\cong} \{0,1\}^X$ that we mentioned just before Definition 4.1.1. In analogy with item (3) we sometimes call a state of the form $unit(x) = 1|x\rangle$ a *point state*.

We now come to the basic definition of validity.

Definition 4.1.2. Let X be a set with an observable $p\colon X \to \mathbb{R}$ on X.

1 For a distribution/state $\omega = \sum_i r_i|x_i\rangle$ on X we define its validity $\omega \models p$ as:

$$\omega \models p \coloneqq \sum\nolimits_i r_i \cdot p(x_i) = \sum_{x \in X} \omega(x) \cdot p(x). \tag{4.2}$$

2 If X is a finite set, with size $|X| \in \mathbb{N}$, one can define the *average* $avg(p)$ of the observable p as its validity in the uniform state $unif_X$ on X, *i.e.*,

$$avg(p) \coloneqq unif_X \models p = \sum_{x \in X} \frac{p(x)}{|X|}.$$

Notice that validity $\omega \models p$ involves a finite sum, since a distribution $\omega \in \mathcal{D}(X)$ has by definition finite support. The sample space X may well be infinite. Validity is often called *expected value* and written as $E[p]$, where the state ω is left implicit. The notation $\omega \models p$ makes the state explicit, which is important in situations where we change the state, for instance via state transformation.

The validity $\omega \models p$ is non-negative (is in $\mathbb{R}_{\geq 0}$) if p is a factor and lies in the unit interval $[0, 1]$ if p is a predicate (whether sharp or not). The fact that the multiplicities r_i in a distribution $\omega = \sum_i r_i|x_i\rangle$ add up to one means that the validity $\omega \models \mathbf{1}$ of truth is one. Notice that for a point predicate one has $\omega \models \mathbf{1}_x = \omega(x)$ and similarly, for a point state, $1|x\rangle \models p = p(x)$.

For a fixed state $\omega \in \mathcal{D}(X)$ we can view $\omega \models (-)$ as a function $Pred(X) \to [0, 1]$ that assigns a likelihood to a belief function (predicate). This is at the heart of the Bayesian interpretation of probability, see Section 6.8 for more details.

As an aside, we typically do not write brackets in equations like $(\omega \models p) = a$, but use the convention that $\models$ has higher precedence than $=$, so that the equation can simply be written as $\omega \models p = a$. Similarly, one can have validity $c \gg\!= \omega \models p$ in a transformed state, which should be read as $(c \gg\!= \omega) \models p$.

Definition 4.1.3. In the presence of a map $incl_X\colon X \hookrightarrow \mathbb{R}$, one can define the mean $mean(\omega)$, also known as average, of a distribution ω on X as the validity of $incl_X$, considered as an observable:

$$mean(\omega) \coloneqq \omega \models incl_X = \sum_{x \in X} \omega(x) \cdot x.$$

We will use the same definition of mean when ω is a multiset instead of a distribution. In fact, we have already been using that extension to multisets in Definition 3.9.1, where we defined Ewens multisets.

Example 4.1.4. 1 Let $flip(\frac{3}{10}) = \frac{3}{10}|1\rangle + \frac{7}{10}|0\rangle$ be a biased coin. Suppose there

is a game where you can throw the coin and win €100 if head (1) comes up, but you lose €50 if the outcome is tail (0). Is it a good idea to play the game?

The possible gain can be formalised as an observable $v\colon \{0,1\} \to \mathbb{R}$ with $v(0) = -50$ and $v(1) = 100$. We get an anwer to the above question by computing the validity:

$$\begin{aligned} \mathit{flip}(\tfrac{3}{10}) \models v &= \sum_{x\in\{0,1\}} \mathit{flip}(\frac{3}{10})(x) \cdot v(x) \\ &= \mathit{flip}(\tfrac{3}{10})(0) \cdot v(0) + \mathit{flip}(\tfrac{3}{10})(1) \cdot v(1) \\ &= \tfrac{7}{10} \cdot -50 + \tfrac{3}{10} \cdot 100 \;=\; -35 + 30 \;=\; -5. \end{aligned}$$

Hence it is wiser not to play.

2 Write *pips* for the set $\{1,2,3,4,5,6\}$, considered as a subset of $\mathbb{R}$, via the map $\mathit{incl}\colon \mathit{pips} \hookrightarrow \mathbb{R}$, which is used as an observable. The average of the latter is its validity $\mathit{dice} \models \mathit{incl}$ for the (uniform) fair dice state $\mathit{dice} = \mathit{unif}_{\mathit{pips}} = \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$. This average is $\frac{21}{6} = \frac{7}{2}$. It is the expected outcome for throwing a dice.

3 Suppose that we claim that in a throw of a (fair) dice the outcome is even. How likely is this claim? We formalise it as a (sharp) predicate $e\colon \mathit{pips} \to [0,1]$ with $e(1) = e(3) = e(5) = 0$ and $e(2) = e(4) = e(6) = 1$. Then, as expected:

$$\begin{aligned} \mathit{dice} \models e &= \sum_{x\in\mathit{pips}} \mathit{dice}(x) \cdot e(x) \\ &= \tfrac{1}{6} \cdot 0 + \tfrac{1}{6} \cdot 1 + \tfrac{1}{6} \cdot 0 + \tfrac{1}{6} \cdot 1 + \tfrac{1}{6} \cdot 0 + \tfrac{1}{6} \cdot 1 \;=\; \tfrac{1}{2}. \end{aligned}$$

Now consider a more subtle claim that the even pips are more likely than the odd ones, where the precise likelihoods are described by the (non-sharp, fuzzy) predicate $p\colon \mathit{pips} \to [0,1]$ with:

$$\begin{array}{lll} p(1) = \frac{1}{10} & p(2) = \frac{9}{10} & p(3) = \frac{3}{10} \\ p(4) = \frac{8}{10} & p(5) = \frac{2}{10} & p(6) = \frac{7}{10}. \end{array}$$

This new claim p happens to be equally probable as the even claim e, since:

$$\begin{aligned} \mathit{dice} \models p &= \tfrac{1}{6} \cdot \tfrac{1}{10} + \tfrac{1}{6} \cdot \tfrac{9}{10} + \tfrac{1}{6} \cdot \tfrac{8}{10} + \tfrac{1}{6} \cdot \tfrac{8}{10} + \tfrac{1}{6} \cdot \tfrac{2}{10} + \tfrac{1}{6} \cdot \tfrac{7}{10} \\ &= \tfrac{1+9+3+8+2+7}{60} \;=\; \tfrac{30}{60} \;=\; \tfrac{1}{2}. \end{aligned}$$

4 Recall the binomial distribution $\mathit{bn}[K](r)$ on the set $\{0,1,\ldots,K\}$ from Example 2.1.1 (2), for $r \in [0,1]$. There is an inclusion function $\{0,1,\ldots,K\} \hookrightarrow \mathbb{R}$ that allows us to compute the mean of the binomial distribution. In this computation we treat the binomial distribution as a special instance of the

multinomial distribution, via the isomorphism $\{0, 1, \ldots, K\} \cong \mathcal{N}[K](2)$. This allows us to re-use an earlier result.

$$\begin{aligned}
&\mathit{mean}(\mathit{bn}[K](r)) \\
&= \sum_{0 \leq k \leq K} \mathit{bn}[K](r)(k) \cdot k \\
&= \sum_{\varphi \in \mathcal{N}[K](2)} \mathit{mn}[K](\mathit{flip}(r))(\varphi) \cdot \varphi(1) \qquad \text{for } \mathit{flip}(r) = r|1\rangle + (1-r)|0\rangle \\
&= K \cdot \mathit{flip}(r)(1) \qquad \text{by Lemma 3.3.4} \\
&= K \cdot r.
\end{aligned}$$

In the two plots of binomial distributions $\mathit{bn}[10](\frac{1}{3})$ and $\mathit{bn}[10](\frac{3}{4})$ in Figure 2.2 (on page 75) one can see that the associated means $\frac{10}{3} = 3.333\ldots$ and $\frac{30}{4} = 7.5$ make sense.

Means for multivariate drawing will be considered in Section 4.4.

For the next result we recall from Proposition 2.4.5 that if M is a commutative monoid, so is the set $\mathcal{D}(M)$ of distributions on M. This is used below where M is the additive monoid $\mathbb{N}$ of natural numbers. The result can be generalised to any additive submonoid of the reals.

Lemma 4.1.5. *The mean function, for distributions on natural numbers, is a homomorphism of monoids of the form:*

$$(\mathcal{D}(\mathbb{N}), +, 0) \xrightarrow{\mathit{mean}} (\mathbb{R}_{\geq 0}, +, 0).$$

Proof. Preservation of the zero element is easy, since:

$$\mathit{mean}(1|0\rangle) = 1 \cdot 0 = 0.$$

Next, for $\omega, \rho \in \mathcal{D}(\mathbb{N})$,

$$\begin{aligned}
\mathit{mean}(\omega + \rho) &\overset{(2.18)}{=} \sum_{n \in \mathbb{N}} \mathcal{D}(+)(\omega \otimes \rho)(n) \cdot n \\
&= \sum_{m,k \in \mathbb{N}} (\omega \otimes \rho)(m,k) \cdot (m+k) \\
&= \sum_{m,k \in \mathbb{N}} \omega(m) \cdot \rho(k) \cdot (m+k) \\
&= \left(\sum_{m,k \in \mathbb{N}} \omega(m) \cdot \rho(k) \cdot m \right) + \left(\sum_{m,k \in \mathbb{N}} \omega(m) \cdot \rho(k) \cdot k \right) \\
&= \left(\sum_{m \in \mathbb{N}} \omega(m) \cdot \left(\sum_{k \in \mathbb{N}} \rho(k) \right) \cdot m \right) + \left(\sum_{k \in \mathbb{N}} \left(\sum_{m \in \mathbb{N}} \omega(m) \right) \cdot \rho(k) \cdot k \right) \\
&= \left(\sum_{m \in \mathbb{N}} \omega(m) \cdot m \right) + \left(\sum_{k \in \mathbb{N}} \rho(k) \cdot k \right) \\
&= \mathit{mean}(\omega) + \mathit{mean}(\rho).
\end{aligned}$$

❑

Remark 4.1.6. What is the difference between a multiset and a factor, and between a distribution and a predicate? A multiset on X is a function $X \to \mathbb{R}_{\geq 0}$ with finite support, and thus a factor. In the same way a distribution on X is a function $X \to [0,1]$ with finite support, and thus a predicate on X. Hence, there are inclusions:

$$\mathcal{M}(X) \subseteq \mathit{Fact}(X) = (\mathbb{R}_{\geq 0})^X \qquad \text{and} \qquad \mathcal{D}(X) \subseteq \mathit{Pred}(X) = [0,1]^X.$$

The first inclusion $\subseteq$ may actually be seen as an equality $=$ when the set X is a finite set. A predicate p on a finite set however, need not be a state, since its values $p(x)$ need not add up to one. However, such a predicate can be normalised; then it is turned into a state.

We are however reluctant to identify states/distributions with certain predicates, since they belong to different universes and have quite different algebraic properties. For instance, distributions are convex sets, whereas predicates are effect modules carrying a commutative monoid, see the next section. Keeping states and predicates apart is a matter of mathematicaly hygiene[1]. We have already seen state transformation $\gg\!=$ along a channel; it preserves the convex structure. Later on in this chapter we shall also see predicate transformation $=\!\ll$ along a channel, in opposite direction; this operation preserves the effect module structure on predicates.

Apart from these mathematical differences, states and predicates play entirely different roles and are understood in different ways: states play an ono-

[1] Moreover, in continuous probability there are no inclusions of states in predicates.

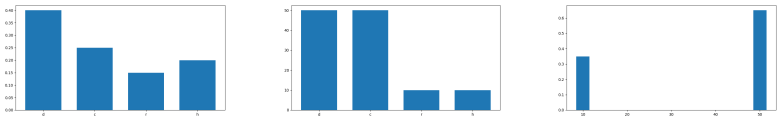
tologoical role and describe 'states of affairs'; predicates are epistemological in nature, and describe evidence (what we know).

Whenever we do make use of the above inclusions, we shall make this usage explicit.

4.1.1 Random variables

So far we have avoided discussing the concept of 'random variable'. It can be a confusing notion for people who start studying probability theory. One often encounters phrases like: let Y be a random variable with expectation $E[Y]$. How should this be understood? For clarity, we define a random variable to be a *pair*, consisting of a distribution and an observable, with a shared underlying space. For such a pair it makes sense to talk about expectation, namely as their validity.

Consider a space $Pet = \{d, c, r, h\}$ for d = dog, c = cat, r = rabbit and h = hamster. We look at the cost (*e.g.* of food) of a pet per month, via an observable $q\colon Pet \to \mathbb{R}$ given by $q(d) = q(c) = 50$ and $q(r) = q(h) = 10$. Let us assume that the distribution of pets in a certain neighbourhood is given by $\omega = \frac{2}{5}|D\rangle + \frac{1}{4}|C\rangle + \frac{3}{20}|R\rangle + \frac{1}{5}|H\rangle$. We can describe the situation via three plots:



The pet distribution ω is on the left and the costs per pet is in the middle. The plot on the right describes $\frac{7}{20}|10\rangle + \frac{13}{20}|50\rangle$, which is the distribution of pet costs in the neighbourhood. It is obtained via the functoriality of $\mathcal{D}$, as image distribution $\mathcal{D}(q)(\omega) \in \mathcal{D}(\mathbb{R})$. In more traditional notation it is described as:

$$P[q = 10] = \tfrac{7}{20} \qquad\qquad P[q = 50] = \tfrac{13}{20}.$$

Sometimes, a random variable is described as such an (image) distribution on the reals, with the underlying distribution ω and observable q left implicit. Alternatively, a rondom variable is sometimes described via a tilde $\sim$, as: $q \sim \omega$, like in phrases such as: "let $q \sim pois[\lambda]$ with ...". This means that $q\colon \mathbb{N} \to \mathbb{R}$ is an observable, on the underlying space $\mathbb{N}$ of the Poisson distribution. It comes close to what our convention will be, namely to describe a random variable as a pair of a state and an observable, with the same underlying space. It should solve what is called in [103, §16] the "notational confusion between a random variable and its distribution".

Definition 4.1.7. 1 A *random variable* on a sample space (set) X consists of two parts:

- a state $\omega \in \mathcal{D}(X)$;
- an observable $R\colon X \to \mathbb{R}$.

2 The probability mass function $P[R = (-)]\colon \mathbb{R} \to [0, 1]$ associated with the random variable (ω, R) is the image distribution on $\mathbb{R}$ given by:

$$P[R = (-)] \coloneqq \mathcal{D}(R)(\omega) = R \gg\!= \omega, \tag{4.3}$$

where R is understood as a deterministic channel in the last expression $R \gg\!= \omega$, see Lemma 1.8.3 (4).

3 The expected value $E[R]$ of a random variable (ω, R) is the validity of R in ω, which can be expressed as mean of the image distribution:

$$\omega \models R = \mathit{mean}(R \gg\!= \omega) = \mathit{mean}(\mathcal{D}(R)(\omega)). \tag{4.4}$$

The image distribution in (4.3) can be described in several (equivalent) ways:

$$\begin{aligned} P[R = r] = \mathcal{D}(R)(\omega)(r) = (R \gg\!= \omega)(r) &= \omega \gg\!= R \models \mathbf{1}_r \\ &= \sum_{x,\, R(x)=r} \omega(x) \\ &= \sum_{x \in R^{-1}(r)} \omega(x) \\ &= \omega \models \mathbf{1}_{R^{-1}(r)}. \end{aligned}$$

In the second item of the above definition, Equation (4.4) holds since:

$$\begin{aligned} \mathit{mean}(R \gg\!= \omega) &= \sum_{r \in \mathbb{R}} (R \gg\!= \omega)(r) \cdot r \qquad \text{see Definition 4.1.3} \\ &= \sum_{r \in \mathbb{R}} \mathcal{D}(R)(\omega)(r) \cdot r \\ &= \sum_{r \in \mathbb{R}} \left(\sum_{x \in R^{-1}(r)} \omega(x) \right) \cdot r \\ &= \sum_{x \in X} \omega(x) \cdot R(x) \\ &= \omega \models R. \end{aligned}$$

Example 4.1.8. We consider the expected value for the sum of two dices. In this situation we have an observable $S : \mathit{pips} \times \mathit{pips} \to \mathbb{R}$, on the sample space $\mathit{pips} = \{1, 2, 3, 4, 5, 6\}$, given by $S(i, j) = i + j$. It forms a random variable together with the product state $\mathit{dice} \otimes \mathit{dice} \in \mathcal{D}(\mathit{pips} \times \mathit{pips})$. Recall,

$\mathit{dice} = \mathit{unif} = \sum_{i \in \mathit{pips}} \frac{1}{6}|i\rangle$ is the uniform distribution unif on pips. First, the distribution for the sum of the pips is the image distribution:

$$\begin{aligned} S \gg\!= \mathit{dice} \otimes \mathit{dice} &= \mathcal{D}(+)(\mathit{dice} \otimes \mathit{dice}) \\ &= \sum_{2 \leq n \leq 12} \Big(\sum_{i+j=n} (\mathit{dice} \otimes \mathit{dice})(i,j) \Big) \big| n \big\rangle \\ &= \tfrac{1}{36}|2\rangle + \tfrac{1}{18}|3\rangle + \tfrac{1}{12}|4\rangle + \tfrac{1}{9}|5\rangle + \tfrac{5}{36}|6\rangle + \tfrac{1}{6}|7\rangle \\ &\qquad + \tfrac{5}{36}|8\rangle + \tfrac{1}{9}|9\rangle + \tfrac{1}{12}|10\rangle + \tfrac{1}{18}|11\rangle + \tfrac{1}{36}|12\rangle. \end{aligned} \tag{4.5}$$

The expected value of the random variable $(\mathit{dice} \otimes \mathit{dice}, S)$ is, according to Definition 4.1.7 (3),

$$\begin{aligned} \mathit{mean}(S \gg\!= \mathit{dice} \otimes \mathit{dice}) &= \mathit{dice} \otimes \mathit{dice} \models S \\ &= \sum_{i,j \in \mathit{pips}} (\mathit{dice} \otimes \mathit{dice})(i,j) \cdot S(i,j) \\ &= \sum_{i,j \in \mathit{pips}} \frac{1}{6} \cdot \frac{1}{6} \cdot (i+j) \\ &= \frac{\sum_{i,j \in \mathit{pips}} i+j}{36} = \frac{252}{36} = 7. \end{aligned}$$

There is a more abstract way to look at this example, using Lemma 4.1.5. We have used dice as a distribution on $\mathit{pips} = \{1, \ldots, 6\}$, but we can also see it as a distribution $\mathit{dice} \in \mathcal{D}(\mathbb{N})$ on the natural numbers. The sum of pips that we are interested in can then be described via a sum of distributions $\mathit{dice} + \mathit{dice}$, using the sum $+$ from Proposition 2.4.5. Then, by Lemma 4.1.5, we also get:

$$\mathit{mean}(\mathit{dice} + \mathit{dice}) = \mathit{mean}(\mathit{dice}) + \mathit{mean}(\mathit{dice}) = \tfrac{7}{2} + \tfrac{7}{2} = 7.$$

We conclude with result for which it is relevant to know in which state we are evaluating an observable.

Lemma 4.1.9. *Let $M = (M, +, 0)$ be commutative monoid, with two distributions $\omega, \rho \in \mathcal{D}(M)$, and with an observable $p \colon M \to \mathbb{R}$ that is a map of (additive) monoids. Then:*

$$(\omega + \rho) \models p = (\omega \models p) + (\rho \models p),$$

where the sum of distributions $\omega + \rho$ is defined in (2.18).

This result involves three different random variables, namely:

$$(\omega + \rho,\ p) \qquad (\omega,\ p) \qquad (\rho, p).$$

Proof. By unravelling the definitions:

$$\begin{aligned}
(\omega + \rho) \models p &= \sum_{x \in M} \mathcal{D}(+)(\omega \otimes \rho)(x) \cdot p(x) \\
&= \sum_{x \in M} \sum_{y,z \in M,\, y+z=x} \omega(y) \cdot \rho(z) \cdot p(x) \\
&= \sum_{y,z \in M} \omega(y) \cdot \rho(z) \cdot p(y+z) \\
&= \sum_{y,z \in M} \omega(y) \cdot \rho(z) \cdot (p(y) + p(z)) \\
&= \sum_{y \in M} \omega(y) \cdot \left(\sum_{z \in M} \rho(z) \right) \cdot p(y) + \sum_{z \in M} \left(\sum_{y \in M} \omega(y) \right) \cdot \rho(z) \cdot p(z) \\
&= \sum_{y \in M} \omega(y) \cdot p(y) + \sum_{z \in M} \rho(z) \cdot p(z) \\
&= (\omega \models p) + (\rho \models p).
\end{aligned}$$

❑

Exercises

4.1.1 Check that the average of the set $n + 1 = \{0, 1, \ldots, n\}$, considered as random variable, is $\frac{n}{2}$.

4.1.2 In Example 4.1.8 we have seen that $\mathit{dice} \otimes \mathit{dice} \models S = 7$, for the observable $S : \mathit{pips} \times \mathit{pips} \to \mathbb{R}$ by $S(x, y) = x + y$.

1 Now define $T : \mathit{pips}^3 \to \mathbb{R}$ by $T(x, y, z) = x + y + z$. Prove that $\mathit{pips} \otimes \mathit{pips} \otimes \mathit{pips} \models T = \frac{21}{2}$.

2 Can you generalise and show that summing on pips^n yields validity $\frac{7n}{2}$?

4.1.3 The birthday paradox tells that with at least 23 people in a room there is a more than 50% chance that at least to of them have the same birthday. This is called a 'paradox', because the number 23 looks surprisingly low.

Let us scale this down so that the problem becomes manageable. Suppose there are three people, each with their birthday in the same week.

1 Show that the probability that they all have different birthdays is $\frac{30}{49}$.

2 Conclude that the probability that at least two of the three have the same birthday is $\frac{19}{49}$.

3 Consider the set $\mathit{days} := \{1, 2, 3, 4, 5, 6, 7\}$. The set $\mathcal{N}[3](\mathit{days})$ of multisets of size three contains the possible combinations of the

three birthdays. Define the sharp predicate $p \colon \mathcal{N}[3](\mathit{days}) \to \{0, 1\}$ by $p(\varphi) = 1$ iff $\mathbf{(}\varphi\mathbf{)} \leq 3$. Check that p holds in those cases where at least two birthdays coincide — see also Exercise 1.5.8.

4 Show that the probability $\frac{19}{49}$ of at least two coinciding birthdays can also be obtained via validity, namely as:

$$\mathit{mn}[3](\mathit{unif}_{\mathit{days}}) \models p \;=\; \tfrac{19}{49}.$$

4.1.4 Consider an arbitrary distribution $\omega \in \mathcal{D}(X)$.

1 Check that for a function $f \colon X \to Y$ and an observable $q \in \mathit{Obs}(Y)$,

$$f \gg\!= \omega \models q \;=\; \omega \models q \circ f$$

2 Observe that we get as special case, for an observable $p \colon X \to \mathbb{R}$,

$$\mathit{mean}(p \gg\!= \omega) \;=\; p \gg\!= \omega \models \mathit{id} \;=\; \omega \models p,$$

where the identity function $\mathit{id} : \mathbb{R} \to \mathbb{R}$ is used as observable.

4.1.5 Let $\omega, \rho \in \mathcal{D}(X)$ be given distributions. Prove the following two equations, in the style of Lemma 4.1.5.

1 For observables $p, q \colon X \to \mathbb{R}$, and thus for random variables, (ω, p) and (ρ, q), one has:

$$\mathit{mean}\Big(\mathcal{D}(p)(\omega) + \mathcal{D}(q)(\rho)\Big) \;=\; (\omega \models p) + (\rho \models q)$$

In the notation of Definition 2 we can also write the left-hand side as: $\mathit{mean}(P[p = (-)] + P[q = (-)])$.

2 For channels $c, d \colon X \nrightarrow \mathbb{R}$,

$$\mathit{mean}\Big((c \gg\!= \omega) + (d \gg\!= \rho)\Big) \;=\; (\omega \models \mathit{mean} \circ c) + (\rho \models \mathit{mean} \circ d).$$

4.1.6 Let $\Omega \in \mathcal{D}(\mathcal{D}(X))$ be a distribution of distributions, with an observable $p \colon X \to \mathbb{R}$. We turn p into a 'validity of p' observable $(-) \models p \colon \mathcal{D}(X) \to \mathbb{R}$ on $\mathcal{D}(X)$. Show that:

$$\mathit{flat}(\Omega) \models p \;=\; \Omega \models \Big((-) \models p\Big).$$

4.1.7 We have introduced the mean in Definition 4.1.3 for distributions whose space is a subset of the reals. By definition, these distributions have finite support. But the definition carries over to distributions with infinite support, as described in Remark 2.1.4. Apply this to the Poisson distribution $\mathit{pois}[\lambda] \in \mathcal{D}_\infty(\mathbb{N})$ and show that

$$\mathit{mean}(\mathit{pois}[\lambda]) \;=\; \lambda.$$

4.1.8 Consider the 'mean' operation from Definition 4.1.3 as a function $\mathit{mean}\colon \mathcal{D}(\mathbb{R}) \to \mathbb{R}$. Show that it interacts in the following way with the unit and flatten maps for the distribution monad $\mathcal{D}$, see Section 2.1.

1 $\mathit{mean} \circ \mathit{unit} = \mathit{id}$;
2 $\mathit{mean} \circ \mathit{flat} = \mathit{mean} \circ \mathcal{D}(\mathit{mean})$.
3 Let $p\colon X \to \mathbb{R}$ be an observable, giving a function $(-) \models p\colon \mathcal{D}(X) \to \mathbb{R}$, sending $\omega \in \mathcal{D}(X)$ to $\omega \models p$ in $\mathbb{R}$. Show that the following diagram commutes.

$$\begin{array}{ccc} \mathcal{D}(\mathcal{D}(X)) & \xrightarrow{\mathcal{D}((-)\models p)} & \mathcal{D}(\mathbb{R}) \\ {\scriptstyle \mathit{flat}}\downarrow & & \downarrow{\scriptstyle \mathit{mean}} \\ \mathcal{D}(X) & \xrightarrow[(-)\models p]{} & \mathbb{R} \end{array}$$

From the first two items we can conclude that *mean* is an (Eilenberg-Moore) algebra of the distribution monad, see Subsection 1.9.4. The third item says that $(-) \models p\colon \mathcal{D}(X) \to \mathbb{R}$ is a homomorphism of algebras.

4.1.9 For a distribution $\omega \in \mathcal{D}(X)$ write $I(\omega)\colon X \to \mathbb{R}_{\geq 0}$ for the *information content* or *surprise* of the distribution ω, defined as factor:

$$I(\omega)(x) \coloneqq \begin{cases} -\log\big(\omega(x)\big) & \text{if } \omega(x) \neq 0, \text{ i.e. if } x \in \mathit{supp}(\omega) \\ 0 & \text{otherwise.} \end{cases}$$

1 Check that Kullback-Leibler divergence, see Definition 2.7.1, can be described as validity: for $\omega, \rho \in \mathcal{D}(X)$ with $\mathit{supp}(\omega) \subseteq \sup(\rho)$,

$$D_{KL}(\omega, \rho) \;=\; \omega \models I(\rho) - I(\omega).$$

2 The *Shannon entropy* $H(\omega)$ of ω and the *cross entropy* $H(\omega, \rho)$ of ω, ρ, are defined as: validities:

$$\begin{aligned} H(\omega) \coloneqq \omega \models I(\omega) \;&=\; -\textstyle\sum_x \omega(x) \cdot \log\big(\omega(x)\big) \\ H(\omega, \rho) \coloneqq \omega \models I(\rho) \;&=\; -\textstyle\sum_x \omega(x) \cdot \log\big(\rho(x)\big) \end{aligned}$$

Thus, Shannon entropy is 'expected surprise'. Show that:

$$H(\omega, \rho) \;=\; H(\omega) + D_{KL}(\omega, \rho).$$

4.1.10 Consider the entropy function H from the previous exercise.

1 Show that $H(\omega) = 0$ implies that $\omega \in \mathcal{D}(X)$ is a point state $\omega = 1|x\rangle$, for a unique element $x \in X$.
2 Show that for $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$,

$$H(\sigma \otimes \tau) \;=\; H(\sigma) + H(\tau).$$

3 Strengthen the previous item to non-entwinedness: for $\omega \in \mathcal{D}(X \times Y)$,

$$H(\omega) = H(\omega[1,0]) + H(\omega[0,1]) \iff \omega = \omega[1,0] \otimes \omega[0,1].$$

Hint: Use for finitely many numbers $r_i, s_i \in [0,1]$, where $\sum_i r_i = 1 = \sum_i s_i$, that $\sum_i r_i \cdot \log(r_i) = \sum_i r_i \cdot \log(s_i)$ implies $r_i = s_i$ for each i.

4.1.11 Let $\omega \in \mathcal{D}(X)$ be an arbitrary state on a finite set X. Use Proposition 2.7.4 (1) to prove that uniform states have maximal Shannon entropy, that is, $H(\omega) \leq H(\mathit{unif})$, where $\mathit{unif} \in \mathcal{D}(X)$ is the uniform distribution.

4.2 The structure of observables

This section describes the algebraic structures that the various types of observables have, without going too much into mathematical details. We don't wish to turn this section into a sequence of formal definitions. Hence we present the essentials and refer to the literature for details. We include the basic fact that mapping a set X to observables on X is functorial, in a suitable sense. This will give rise to the notion of *weakening*. It plays the same (or dual) role for observables that marginalisation plays for states.

It turns out that our four types of observables are all (commutative) monoids, via multiplication / conjunction, but in different universes. The findings are summarised in the following table.

name	type	monoid in
observable	$X \to \mathbb{R}$	ordered vector spaces
factor	$X \to \mathbb{R}_{\geq 0}$	ordered cones
predicate	$X \to [0,1]$	effect modules
sharp predicate	$X \to \{0,1\}$	join semilattices

(4.6)

The least well-known structures in this table are effect modules. They will thus be described in greatest detail, in Subsection 4.2.3.

4.2.1 Observables

Observables are $\mathbb{R}$-valued functions on a set (or sample space) which are often written as capitals $X, Y, \ldots$. Here, these letters are typically used for sets and spaces. We shall use letters $p, q, \ldots$ for observables in general, and in particular for predicates. In some settings one allows observables $X \to \mathbb{R}^n$ to the n-dimensional space of real numbers. Whenever needed, we shall use such maps as n-ary tuples $\langle p_1, \ldots, p_n \rangle \colon X \to \mathbb{R}^n$ of observables $p_i \colon X \to \mathbb{R}$, see also Section 1.1.

Let us fix a set X, and consider the collection $\mathcal{O}bs(X) = \mathbb{R}^X$ of observables on X. What structure does it have?

- Given two observables $p, q \in \mathcal{O}bs(X)$, we can add them pointwise, giving $p + q \in \mathcal{O}bs(X)$ via $(p + q)(x) = p(x) + q(x)$.
- Given an observable $p \in \mathcal{O}bs(X)$ and a scalar $s \in \mathbb{R}$ we can form a 'rescaled' observable $s \cdot p \in \mathcal{O}bs(X)$ via $(s \cdot p)(x) = s \cdot p(x)$. In this way we get $-p = (-1) \cdot p$ and $\mathbf{0} = 0 \cdot p$ for any observable p, where $\mathbf{0} = \mathbf{1}_\emptyset \in \mathcal{O}bs(X)$ is taken from Definition 4.1.1 (2).
- For observables $p, q \in \mathbb{R}^X$ we have a partial order $p \leq q$ defined by: $p(x) \leq q(x)$ for all $x \in X$.

Together these operations of sum $+$ and scalar multiplication $\cdot$ make $\mathbb{R}^X$ into a *vector space* over the real numbers, since $+$ and $\cdot$ satisfy the appropriate axioms of vector spaces. Moreover, this is an *ordered* vector space by the third bullet.

One can restricts to *bounded* observables $p \colon X \to \mathbb{R}$ for which there is a bound $B \in \mathbb{R}_{>0}$ such that $-B \leq p(x) \leq B$ for all $x \in X$. The collection of such bounded observables forms an *order unit space* [3, 83, 127, 132].

On $\mathcal{O}bs(X) = \mathbb{R}^X$ there is also a commutative monoid structure $(\mathbf{1}, \&)$, where $\&$ is pointwise multiplication: $(p \,\&\, q)(x) = p(x) \cdot q(x)$. We prefer to write this operation as logical conjunction because that is what it is when restricted to predicates. Besides, having yet another operation that is written as dot $\cdot$ might be confusing. We will occasionally write p^n for $p \,\&\, \cdots \,\&\, p$ (n times).

4.2.2 Factors

We recall that a factor is a non-negative observables and that we write $\mathcal{F}act(X) = (\mathbb{R}_{\geq 0})^X$ for the set of factors on a set X. Factors occur in probability theory for instance in the so-called sum-product algorithms for computing marginals and posterior distributions in graphical models. These matters are postponed to Chapter **??**; here we concentrate on the mathematical properties of factors.

The set $\mathcal{F}act(X)$ looks like a vector space, except that there are no negatives.

Using the notation introduced in the previous subsection, we have:

$$Fact(X) = \{p \in Obs(X) \mid p \geq \mathbf{0}\}.$$

Factors can be added pointwise, with identity element $\mathbf{0} \in Fact(X)$, like random variables. But a factor $p \in Fact(X)$ cannot be re-scaled with an arbitrary real number, but only with a non-negative number $s \in \mathbb{R}_{\geq 0}$, giving $s \cdot p \in Fact(X)$. These structures are often called *cones*. The cone $Fact(X)$ is *positive*, since $p + q = \mathbf{0}$ implies $p = q = \mathbf{0}$. It is also *cancellative*: $p + r = q + r$ implies $p = q$.

The monoid $(\mathbf{1}, \&)$ on $Obs(X)$ restricts to $Fact(X)$, since $\mathbf{1} \geq \mathbf{0}$ and if $p, q \geq \mathbf{0}$ then also $p \,\&\, q \geq \mathbf{0}$.

4.2.3 Predicates

We first note that the set $Pred(X) = [0,1]^X$ of predicates on a set X contains falsity $\mathbf{0}$ and truth $\mathbf{1}$, which are always 0 (resp. 1). There are some noteworthy differences between predicates on the one hand and observables and factors on the other hand.

- Predicates are not closed under addition, since the sum of two numbers in $[0,1]$ may ly outside $[0,1]$. Thus, addition of predicates is a *partial* operation, and is then written as $p \owedge q$. Thus: $p \owedge q$ is defined if $p(x) + q(x) \leq 1$ for all $x \in X$, and in that case $(p \owedge q)(x) = p(x) + q(x)$.

  This operation $\owedge$ is commutative and associative in a suitably partial sense. Moreover, it has $\mathbf{0}$ as identity element: $p \owedge \mathbf{0} = p = \mathbf{0} \owedge p$. This is structure $(Pred(X), \mathbf{0}, \owedge)$ is called a *partial commutative monoid*.
- There is a 'negation' of predicates, written as $p^\perp$, and called *orthosupplement*. It is defined as $p^\perp = \mathbf{1} - p$, that is, as $p^\perp(x) = 1 - p(x)$. Then: $p \owedge p^\perp = \mathbf{1}$ and $p^{\perp\perp} = p$.
- Predicates are closed under scalar multiplication $s \cdot p$, but only if one restricts the scalar s to be in the unit interval $[0,1]$. Such scalar multiplication interacts nicely with partial addition $\owedge$, in the sense that $s \cdot (p \owedge q) = (s \cdot p) \owedge (s \cdot q)$.

The combination of these items means that the set $Pred(X)$ carries the structure of an *effect module* [68], see also [44]. These structures arose in mathematical physics [49] in order to axiomatise the structure of quantum predicates on Hilbert spaces.

The effect module $Pred(X)$ also carries a commutative monoid structure for conjunction, namely $(\mathbf{1}, \&)$. Indeed, when $p, q \in Pred(X)$, then also $p \,\&\, q \in Pred(X)$. We have $p \,\&\, \mathbf{0} = \mathbf{0}$ and $p \,\&\, (q_1 \owedge q_2) = (p \,\&\, q_1) \owedge (p \,\&\, q_2)$.

Since these effect structures are not so familiar, we include more formal descriptions.

Definition 4.2.1. 1 A *partial commutative monoid* (PCM) consists of a set M with a zero element $0 \in M$ and a partial binary operation $\otimes\colon M \times M \to M$ satisfying the three requirements below. They involve the notation $x \perp y$ for: $x \otimes y$ is defined; in that case x, y are called orthogonal.

- Commutativity: $x \perp y$ implies $y \perp x$ and $x \otimes y = y \otimes x$;
- Associativity: $y \perp z$ and $x \perp (y \otimes z)$ implies $x \perp y$ and $(x \otimes y) \perp z$ and also $x \otimes (y \otimes z) = (x \otimes y) \otimes z$;
- Zero: $0 \perp x$ and $0 \otimes x = x$;

2 An *effect algebra* is a PCM $(E, 0, \otimes)$ with an orthosupplement. The latter is a total unary 'negation' operation $(-)^\perp\colon E \to E$ satisfying:

- $x^\perp \in E$ is the unique element in E with $x \otimes x^\perp = 1$, where $1 = 0^\perp$;
- $x \perp 1 \Rightarrow x = 0$.

A homomorphism $E \to D$ of effect algebras is given by a function $f\colon E \to D$ between the underlying sets satisfying $f(1) = 1$, and if $x \perp x'$ in E then both $f(x) \perp f(x')$ in D and $f(x \otimes x') = f(x) \otimes f(x')$. Effect algebras and their homomorphisms form a category, denoted by **EA**.

3 An *effect module* is an effect algebra E with a scalar multiplication $s \cdot x$, for $s \in [0, 1]$ and $x \in E$ forming an action:

$$1 \cdot x = x \qquad (r \cdot s) \cdot x = r \cdot (s \cdot x),$$

and preserving sums (that exist) in both arguments:

$$0 \cdot x = 0 \qquad (r + s) \cdot x = r \cdot x \otimes s \cdot x$$
$$s \cdot 0 = 0 \qquad s \cdot (x \otimes y) = s \cdot x \otimes s \cdot y.$$

We write **EMod** for the category of effect modules, where morphisms are maps of effect algebras that preserve scalar multiplication (*i.e.* are 'equivariant').

The following notion of 'test' comes from a quantum context and captures 'compatible' observations, see *e.g.* [36, 130]. It will be used occasionally later on, for instance in Exercise 6.1.4 or Proposition 6.7.10.

Definition 4.2.2. A *test* or, more explicitly, an *n-test* on a set X is an n-tuple of predicates $p_1, \ldots, p_n\colon X \to [0, 1]$ satisfying $p_1 \otimes \cdots \otimes p_n = \mathbf{1}$.

This notion of test can be formulated in an arbitrary effect algebra. Here we do it in $\mathcal{P}red(X)$ only.

Each predicate p forms a 2-test $p, p^\perp$. Exercise 4.2.10 asks to show that an n-test of predicates on X corresponds to a channel $X \nrightarrow n$. In particular, each predicate on X corresponds to a channel $X \nrightarrow 2$, see Exercise 4.3.5.

The probabilities $\omega(x_i) \in [0, 1]$ of a distribution form a test on a singleton set 1.

The following easy observations give a normal form for predicates on a finite set.

Lemma 4.2.3. *Consider a finite set $X = \{x_1, \ldots, x_n\}$.*

1 *The point predicates $\mathbf{1}_{x_1}, \ldots, \mathbf{1}_{x_n}$ form a test on X.*
2 *Each predicate $p \colon X \to [0, 1]$ has a normal form $p = \bigotimes_i r_i \cdot \mathbf{1}_{x_i}$, with scalars $r_i = p(x_i) \in [0, 1]$.* ❑

On a set $X = \{a, b, c\}$ we can write a state and a predicate as:

$$\tfrac{1}{3}|a\rangle + \tfrac{1}{6}|b\rangle + \tfrac{1}{2}|c\rangle \qquad\qquad \tfrac{2}{3} \cdot \mathbf{1}_a \otimes \tfrac{5}{6} \cdot \mathbf{1}_b \otimes 1 \cdot \mathbf{1}_c.$$

Writing $\otimes$ looks a bit pedantic, so we often simply write + instead. Recall that in a predicate the probabilities need not add up to one, see Remark 4.1.6.

A set of predicates $p_1, \ldots, p_n$ on the same space, can be turned into a test via (pointwise) normalisation. Below we describe an alternative construction which is sometimes called *stick breaking*, see *e.g.* [94, Defn. 1]. It is commonly used for probabilities $r_1, \ldots, r_n \in [0, 1]$, but it applies to predicates as well — but it may be described even more generally, inside an arbitrary effect algebra with a commutative monoid structure for conjunction. The idea is to break up a stick in two pieces, multiple times, first according to the first ratio r_1. You then put one part, say on the left, aside and break up the other part according to ratio r_2. Etcetera.

Lemma 4.2.4. *Let $p_1, \ldots, p_n$ be arbitrary predicates, all on the same set, for $n \geq 1$. We can turn them via "stick breaking" into an $n+1$-test $q_1, \ldots, q_{n+1}$ via definitions:*

$$\begin{aligned} q_1 &\coloneqq p_1 \\ q_{i+1} &\coloneqq p_1^\perp \mathbin{\&} \cdots \mathbin{\&} p_i^\perp \mathbin{\&} p_{i+1} \qquad \text{for } 0 < i < n \\ q_{n+1} &\coloneqq p_1^\perp \mathbin{\&} \cdots \mathbin{\&} p_n^\perp. \end{aligned}$$

Proof. We prove $\bigotimes_{i \leq n+1} q_i = \mathbf{1}$ by induction on $n \geq 1$. For $n = 1$ we get $q_{n+1} = q_2 = p_1^\perp$, and then clearly: $q_1 \otimes q_2 = p_1 \otimes p_1^\perp = \mathbf{1}$. Next, using that

multiplication & distributes over sum $\otimes$, we get:

$$\begin{aligned}
& q_1 \otimes q_2 \otimes \cdots \otimes q_{n+1} \otimes q_{n+2} \\
&= p_1 \otimes (p_1^\perp \,\&\, p_2) \otimes (p_1^\perp \,\&\, p_2^\perp \,\&\, p_3) \otimes \cdots \\
&\qquad \cdots \otimes (p_1^\perp \,\&\, p_2^\perp \,\&\, \cdots \,\&\, p_{n-1}^\perp \,\&\, p_n) \otimes (p_1^\perp \,\&\, \cdots \,\&\, p_{n+1}^\perp) \\
&= p_1 \otimes \Big(p_1^\perp \,\&\, \Big[p_2 \otimes (p_2^\perp \,\&\, p_3) \otimes \cdots \\
&\qquad \cdots \otimes (p_2^\perp \,\&\, p_3^\perp \,\&\, \cdots \,\&\, p_{n-1}^\perp \,\&\, p_n) \otimes (p_2^\perp \,\&\, \cdots \,\&\, p_{n+1}^\perp)\Big]\Big) \\
&\overset{\text{(IH)}}{=} p_1 \otimes (p_1^\perp \,\&\, \mathbf{1}) \\
&= \mathbf{1}.
\end{aligned}$$

❑

As mentioned above, the stick breaking construction is commonly used for probabilities $r_i \in [0, 1]$. That is a special case of the above construction, with singleton sample space. We can reformulation this construction as a stick breaking channel *stbr* of the form:

$$[0, 1]^n \xrightarrow{\;stbr\;} \mathcal{D}(n+1) = \mathcal{D}(\{0, 1, \ldots, n\}),$$

with:

$$\begin{aligned}
stbr(r_1, \ldots, r_n) \coloneqq\; & r_1 \big|0\big\rangle + (1-r_1)r_2 \big|1\big\rangle + \cdots + \big(\textstyle\prod_{i<n}(1-r_i)\big)r_n \big|n-1\big\rangle \\
& + \big(\textstyle\prod_{i<n+1}(1-r_i)\big)\big|n\big\rangle.
\end{aligned} \tag{4.7}$$

4.2.4 Sharp predicates

The structure of the set $SPred(X) = \{0, 1\}^X$ is most familiar to logicians: it forms a Boolean algebra. The join $p \vee q$ of $p, q \in SPred(X)$ is the pointwise join $(p \vee q)(x) = p(x) \vee q(x)$, where the latter disjunction $\vee$ is the obvious one on $\{0, 1\}$. One can see these disjunctions $(\mathbf{0}, \vee)$ as forming an additive structure with negation $(-)^\perp$. Conjunction $(\mathbf{1}, \&)$ forms a commutative structure on top.

Formally one can say that $SPred(X)$ also has scalar multiplication, with scalars from $\{0, 1\}$, in such a way that $0 \cdot p = \mathbf{0}$ and $1 \cdot p = p$.

Exercise 4.2.9 below contains several alternative characterisations of sharpness for predicates.

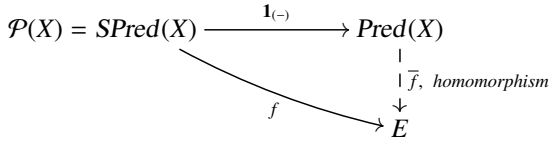
The idea of a free structure as a 'minimal extension' has occurred earlier *e.g.* from Proposition 1.2.3 or 1.4.4. It also applies in the context of predicates, where fuzzy predicates are a free extension of sharp predicates. This is a fundamental insight that will be formulated in terms of effect algebras and modules. Later on in Subsection **??** we will see that such a free extension also exists in continuous probability and forms the essence of integration (as validity).

Theorem 4.2.5. *1 For an arbitrary set X, the indicator function*

$$\mathcal{P}(X) = SPred(X) \xrightarrow{\mathbf{1}_{(-)}} Pred(X)$$

is a homomorphism of effect algebras.

2 Let X now be finite. This indicator map makes Pred(X) *into the* free effect module *on* $\mathcal{P}(X)$*, as described below: for each effect module E with a map of effect algebras* $f\colon \mathcal{P}(X) \to E$*, there is a unique map of effect modules* $\overline{f}\colon Pred(X) \to E$ *in:*



Proof. 1 See Exercise 4.2.3.

2 We use that each predicate $p \in Pred(X)$ can be written in normal form as $p = \bigotimes_{x\in X} p(x) \cdot \mathbf{1}_x$, see Lemma 4.2.3 (2). So we define $\overline{f}$ as:

$$\overline{f}(p) \coloneqq \bigotimes_{x\in X} p(x) \cdot f(\{x\}).$$

Obviously, $\overline{f}(\mathbf{0}) = \mathbf{0}$. Also:

$$\overline{f}(\mathbf{1}) = \bigotimes_{x\in X} 1 \cdot f(\{x\}) = f\left(\bigcup_{x\in X} \{x\}\right) = f(X) = 1.$$

If $p \perp q$, then $p(x) + q(x) \leq 1$ for each $x \in X$. Hence:

$$\begin{aligned}
\overline{f}(p \otimes q) &= \bigotimes_{x\in X} (p \otimes q)(x) \cdot f(\{x\}) = \bigotimes_{x\in X} (p(x) + q(x)) \cdot f(\{x\}) \\
&= \bigotimes_{x\in X} p(x) \cdot f(\{x\}) \otimes q(x) \cdot f(\{x\}) \\
&= \bigotimes_{x\in X} p(x) \cdot f(\{x\}) \otimes \bigotimes_{x\in X} q(x) \cdot f(\{x\}) \\
&= \overline{f}(p) \otimes \overline{f}(q).
\end{aligned}$$

Scalar multiplication is also preserved:

$$\begin{aligned}
\overline{f}(r \cdot p) &= \bigotimes_{x\in X} (r \cdot p)(x) \cdot f(\{x\}) = \bigotimes_{x\in X} r \cdot (p(x) \cdot f(\{x\})) \\
&= r \cdot \bigotimes_{x\in X} p(x) \cdot f(\{x\}) \\
&= r \cdot \overline{f}(p).
\end{aligned}$$

Finally, for uniqueness, let $g\colon \mathit{Pred}(X) \to E$ be a map of effect modules with $g(\mathbf{1}_U) = f(U)$ for each $U \in \mathcal{P}(X)$. Then $g = \overline{f}$ since:

$$\begin{aligned}\overline{f}(p) \;=\; \bigotimes_{x\in X} p(x)\cdot f(\{x\}) \;&=\; \bigotimes_{x\in X} p(x)\cdot g(\mathbf{1}_x) \\ &= g\left(\bigotimes_{x\in X} p(x)\cdot \mathbf{1}_x\right) \;=\; g(p). \qquad \blacksquare\end{aligned}$$

Now that we have seen observables, with factors and (sharp) predicates as special cases, we can say that all these subsets of observables all share the same multiplicative structure $(\mathbf{1}, \&)$ for conjunction, but that their additive structures and scalar multiplications differ. The latter, however, are preserved under taking validity — and not $(\mathbf{1}, \&)$.

Lemma 4.2.6. *Let $\omega \in \mathcal{D}(X)$ be a distribution on a set X. Operations on observables p, q on X satisfy, whenever defined,*

1 $\omega \models \mathbf{0} = 0$*;*
2 $\omega \models (p + q) = (\omega \models p) + (\omega \models q)$*;*
3 $\omega \models p^{\perp} = 1 - (\omega \models p)$*;*
4 $\omega \models (s \cdot p) = s \cdot (\omega \models p)$*.* $\blacksquare$

4.2.5 Parallel products and weakening

Earlier we have seen parallel products $\otimes$ of states and of channels. This $\otimes$ can be defined for observables too, and is then often called parallel conjunction. The difference between parallel conjunction $\otimes$ and sequential conjunction $\&$ is that $\otimes$ acts on observables on *different* sets X, Y and yields an outcome on the product set $X \times Y$, whereas $\&$ works for observables on the same set Z, and produces a conjunction observable again on Z. These $\otimes$ and $\&$ are interdefinable, via transformation $\mathrel{=\!\!\ll}$ of observables, see Section 4.3 — in particular Exercise 4.3.7.

Definition 4.2.7. 1 Let p be an observable on a set X, and q on Y. Then we define a new observable $p \otimes q$ on $X \times Y$ by:

$$(p \otimes q)(x, y) \;\coloneqq\; p(x) \cdot q(y).$$

2 Suppose we have an observable p on a set X and we like to use p on the product $X \times Y$. This can be done by taking $p \otimes \mathbf{1}$ instead. This $p \otimes \mathbf{1}$ is called a *weakening* of p. It satisfies $(p \otimes \mathbf{1})(x, y) = p(x)$.

More generally, consider a product $X_1 \times \cdots \times X_n$. For an observable p on

the i-th set X_i, we weaken p to a predicate on the whole product $X_1 \times \cdots \times X_n$, namely:

$$\underbrace{\mathbf{1} \otimes \cdots \otimes \mathbf{1}}_{i-1 \text{ times}} \otimes p \otimes \underbrace{\mathbf{1} \otimes \cdots \otimes \mathbf{1}}_{n-i \text{ times}}.$$

Weakening is a structural operation in logic which makes it possible to use a predicate $p(x)$ depending on a single variable x in a larger context where one has for instance two variables x, y by ignoring the additional variable y. Weakening is usually not an explicit operation, except in settings like linear logic where one has to be careful about the use of resources. Here, we need weakening as an explicit operation in order to avoid type mismatches between observables and underlying sets.

Recall that marginalisation of states is an operation that moves a state to a smaller underlying set by projecting away. Weakening can be seen as a dual operation, moving an observable to a larger context. There is a close relationship between marginalisation and weakening via validity: for a state $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n)$ and an observable p on X_i we have:

$$\begin{array}{rcl} \omega \models \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes p \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} & = & \mathcal{D}(\pi_i)(\omega) \models p \\ & = & \omega[0, \ldots, 0, 1, 0, \ldots, 0] \models p, \end{array} \tag{4.8}$$

where the 1 in the latter marginalisation mask is at position i. Soon we shall see an alternative description of weakening in terms of predicate transformation. The above equation then appears as a special case of a more general result, namely of Proposition 4.3.3.

We mention some basic results about parallel products of observables. More such 'logical' results can be found in the exercises.

Lemma 4.2.8. *For states $\sigma \in \mathcal{D}(X)$, $\tau \in \mathcal{D}(Y)$ and observables $p \in Obs(X)$, $q \in Obs(Y)$ one has:*

$$(\sigma \otimes \tau \models p \otimes q) = (\sigma \models p) \cdot (\tau \models q).$$

Proof. Easy:

$$\begin{aligned}(\sigma \otimes \tau \models p \otimes q) &= \sum_{z \in X \times Y} (\sigma \otimes \tau)(z) \cdot (p \otimes q)(z) \\ &= \sum_{x \in X, y \in Y} (\sigma \otimes \tau)(x, y) \cdot (p \otimes q)(x, y) \\ &= \sum_{x \in X, y \in Y} \sigma(x) \cdot \tau(y) \cdot p(x) \cdot q(y) \\ &= \left(\sum_{x \in X} \sigma(x) \cdot p(x) \right) \cdot \left(\sum_{y \in Y} \tau(y) \cdot q(y) \right) \\ &= (\sigma \models p) \cdot (\tau \models q). \end{aligned}$$

❑

Lemma 4.2.9. *1 For observables $p_i, q_i \in \mathcal{O}bs(X_i)$ one has:*

$$(p_1 \otimes \cdots \otimes p_n) \,\&\, (q_1 \otimes \cdots \otimes q_n) = (p_1 \,\&\, q_1) \otimes \cdots \otimes (p_n \,\&\, q_n).$$

2 Parallel composition $p \otimes q$ of observables $p \in \mathcal{O}bs(X)$ and $q \in \mathcal{O}bs(Y)$ can be defined in terms of weakening and conjunction, namely as:

$$p \otimes q = (p \otimes \mathbf{1}) \,\&\, (\mathbf{1} \otimes q).$$

Proof. 1 For elements $x_i \in X_i$,

$$\begin{aligned}&\Big((p_1 \otimes \cdots \otimes p_n) \,\&\, (q_1 \otimes \cdots \otimes q_n)\Big)(x_1, \ldots, x_n) \\ &= (p_1 \otimes \cdots \otimes p_n)(x_1, \ldots, x_n) \cdot (q_1 \otimes \cdots \otimes q_n)(x_1, \ldots, x_n) \\ &= (p_1(x_1) \cdot \ldots \cdot p_n(x_n)) \cdot (q_1(x_1) \cdot \ldots \cdot q_n(x_n)) \\ &= (p_1(x_1) \cdot q_1(x_1)) \cdot \ldots \cdot (p_n(x_n) \cdot q_n(x_n)) \\ &= (p_1 \,\&\, q_1)(x_1) \cdot \ldots \cdot (p_n \,\&\, q_n)(x_n) \\ &= \Big((p_1 \,\&\, q_1) \otimes \cdots \otimes (p_n \,\&\, q_n)\Big)(x_1, \ldots, x_n).\end{aligned}$$

2 Directly from the previous item:

$$(p \otimes \mathbf{1}) \,\&\, (\mathbf{1} \otimes q) = (p \,\&\, \mathbf{1}) \otimes (\mathbf{1} \,\&\, q) = p \otimes q.$$

❑

4.2.6 Functoriality

We like to conclude this section with some categorical observations. They are not immediately relevant for the sequel and may be skipped. We shall be using four (new) categories:

- **Vect**, with vector spaces (over the real numbers) as objects and linear maps as morphisms between them (preserving addition and scalar multiplication);
- **Cone**, with cones as objects and also with linear maps as morphisms, but this time preserving scalar multiplication with non-negative reals only;

- **EMod**, with effect modules as objects and homomorphisms of effect modules as maps between them (see Definition 4.2.1);
- **BA**, with Boolean algebras as objects and homomorphisms of Boolean algebras (preserving finite joins and negations, and then also finite meets).

Recall from Subsection 1.9.1 that we write $\mathbb{C}^{\mathrm{op}}$ for the opposite of category $\mathbb{C}$, with arrows reversed. This opposite is needed in the following result.

Proposition 4.2.10. *Taking particular observables on a set is functorial: there are functors:*

1 $Obs \colon \mathbf{Sets} \to \mathbf{Vect}^{\mathrm{op}}$;
2 $Fact \colon \mathbf{Sets} \to \mathbf{Cone}^{\mathrm{op}}$;
3 $Pred \colon \mathbf{Sets} \to \mathbf{EMod}^{\mathrm{op}}$;
4 $SPred \colon \mathbf{Sets} \to \mathbf{BA}^{\mathrm{op}}$.

On maps $f \colon X \to Y$ in **Sets** *these functors are all defined by the 'pre-compose with f' operation $q \mapsto q \circ f$. They thus reverse the direction of morphisms, which necessitates the use of opposite categories $(-)^{\mathrm{op}}$.*

The above functors all preserve the partial order on observables and also the commutative monoid structure $(\mathbf{1}, \&)$, since they are defined pointwise.

Proof. We consider the first instance of observables in some detail. The other cases are similar. For a set X we have have seen that $Obs(X) = \mathbb{R}^X$ is a vector space, and thus an object of the category **Vect**. Each function $f \colon X \to Y$ in **Sets** gives rise to a function $Obs(f) \colon Obs(Y) \to Obs(X)$ in the opposite direction. It maps an observable $q \colon Y \to \mathbb{R}$ on Y to the observable $q \circ f \colon X \to \mathbb{R}$ on X. It is not hard to see that this function $Obs(f) = (-) \circ f$ preserves the vector space structure. For instance, it preserves sums, since they are defined pointwise. We shall prove this in a precise, formal manner. First $Obs(f)(\mathbf{0})$ is the function that map $x \in X$ to:

$$Obs(f)(\mathbf{0})(x) = (\mathbf{0} \circ f)(x) = \mathbf{0}(f(x)) = 0.$$

Hence $Obs(f)(\mathbf{0})$ maps everything to 0 and is thus equal to the zero function itself: $Obs(f)(\mathbf{0}) = \mathbf{0}$. Next, addition $+$ is preserved since:

$$\begin{aligned} Obs(f)(p+q) &= (p+q) \circ f \\ &= [x \mapsto (p+q)(f(x))] \\ &= [x \mapsto p(f(x)) + q(f(x))] \\ &= [x \mapsto Obs(f)(p)(x) + Obs(f)(q)(x)] \\ &= Obs(f)(p) + Obs(f)(q). \end{aligned}$$

We leave preservation of scalar multiplication to the reader and conclude that

$Obs(f)$ is a linear function, and thus a morphism $Obs(f)\colon Obs(Y) \to Obs(X)$ in **Vect**. Hence $Obs(f)$ is a morphism $Obs(X) \to Obs(Y)$ in the opposite category $\mathbf{Vect}^{\mathrm{op}}$. We still need to check that identity maps and composition is preserved. We do the latter. For $f\colon X \to Y$ and $g\colon Y \to Z$ in **Sets** we have, for $r \in Obs(Z)$,

$$\begin{aligned} Obs(g \circ f)(r) \;=\; r \circ (g \circ f) &= (r \circ g) \circ f \\ &= Obs(g)(r) \circ f \\ &= Obs(f)(Obs(g)(r)) \\ &= (Obs(f) \circ Obs(g))(r) \\ &= (Obs(g) \circ^{\mathrm{op}} Obs(f))(r). \end{aligned}$$

This yields $Obs(g \circ f) = Obs(g) \circ^{\mathrm{op}} Obs(f)$, so that we get a functor of the form $Obs\colon \mathbf{Sets} \to \mathbf{Vect}^{\mathrm{op}}$. ❑

Notice that saying that we have a functor like $Pred\colon \mathbf{Sets} \to \mathbf{EMod}^{\mathrm{op}}$ contains remarkably much information, about the mathematical structure on objects $Pred(X)$, about preservation of this structure by maps $Pred(f)$, and about preservation of identity maps and composition by $Pred(-)$ on morphisms (see also Exercise 4.3.10). This makes the language of category theory both powerful and efficient.

Exercises

4.2.1 Consider the following question:

An urn contains 10 balls of which 4 are red and 6 are blue. A second urn contains 16 red balls and an unknown number of blue balls. A single ball is drawn from each urn. The probability that both balls are the same color is $\frac{11}{25}$. Find the number of blue balls in the second urn.

1 Check that the givens can be expressed in terms of validity as:

$$\tfrac{11}{25} \;=\; Flrn(4|R\rangle{+}6|B\rangle) \otimes Flrn(16|R\rangle{+}x|B\rangle) \models (\mathbf{1}_R \otimes \mathbf{1}_R) \between (\mathbf{1}_B \otimes \mathbf{1}_B)$$

2 Prove, by solving the above equation, that there are 4 blue balls in the second urn.

4.2.2 Find examples of predicates p, q on a set X and a distribution ω on X such that $\omega \models p \mathbin{\&} q$ and $(\omega \models p) \cdot (\omega \models q)$ are different.

4.2.3 1 Check that (sharp) indicator predicates $\mathbf{1}_E\colon X \to [0,1]$, for subsets $E \subseteq X$, satisfy:

- $\mathbf{1}_{E \cap D} = \mathbf{1}_E \mathbin{\&} \mathbf{1}_D$, and thus $\mathbf{1} \otimes \mathbf{1} = \mathbf{1}$, $\mathbf{1} \otimes \mathbf{0} = \mathbf{0} = \mathbf{0} \otimes \mathbf{1}$;
- $\mathbf{1}_{E \cup D} = \mathbf{1}_E \between \mathbf{1}_D$, if E, D are disjoint;

- $(\mathbf{1}_E)^\perp = \mathbf{1}_{\neg E}$, where $\neg E = X \setminus E = \{x \in X \mid x \notin E\}$ is the complement of E.

Formally, the function $\mathbf{1}_{(-)} \colon \mathcal{P}(X) \to \mathit{Pred}(X) = [0,1]^X$ is a homomorphism of effect algebras, see Theorem 4.2.5 (1).

2 Show that this indicator function is natural, in the sense that for $f \colon X \to Y$ the following diagram commutes.

$$\begin{array}{ccc} \mathcal{P}(X) & \xrightarrow{\mathbf{1}_{(-)}} & \mathit{Pred}(X) \\ {\scriptstyle f^{-1}}\uparrow & & \uparrow{\scriptstyle (-)\circ f} \\ \mathcal{P}(Y) & \xrightarrow{\mathbf{1}_{(-)}} & \mathit{Pred}(Y) \end{array}$$

3 Now consider subsets $E \subseteq X$ and $D \subseteq Y$ of different sets X, Y together with their product subset $E \times D \subseteq X \times Y$. Show that $\mathbf{1}_{E\times D} = \mathbf{1}_E \otimes \mathbf{1}_D$, with as special case $\mathbf{1}_{(x,y)} = \mathbf{1}_x \otimes \mathbf{1}_y$.

4.2.4 One may expect the following implication between inequalities of validities:

$$(\sigma \models p) \leq (\tau \models p) \Longrightarrow (\sigma \models p \mathbin{\&} q) \leq (\tau \models p \mathbin{\&} q).$$

However, it *fails*. This exercise elaborates a counterexample. Take $X = \{a, b, c\}$ with states:

$$\sigma = \tfrac{19}{100}|a\rangle + \tfrac{47}{100}|b\rangle + \tfrac{17}{50}|c\rangle \qquad \tau = \tfrac{1}{5}|a\rangle + \tfrac{9}{20}|b\rangle + \tfrac{7}{20}|c\rangle$$

with predicates:

$$p = 1 \cdot \mathbf{1}_a + \tfrac{7}{10} \cdot \mathbf{1}_b + \tfrac{1}{2} \cdot \mathbf{1}_c \qquad q = \tfrac{1}{10} \cdot \mathbf{1}_a + \tfrac{1}{2} \cdot \mathbf{1}_b + \tfrac{1}{5} \cdot \mathbf{1}_c.$$

Now check consecutively that:

1 $\sigma \models p = \frac{689}{1000} < \frac{69}{100} = \tau \models p$.

2 $p \mathbin{\&} q = \frac{1}{10} \cdot \mathbf{1}_a + \frac{7}{20} \cdot \mathbf{1}_b + \frac{1}{10} \cdot \mathbf{1}_c$.

3 $\sigma \models p \mathbin{\&} q = \frac{87}{400} > \frac{85}{400} = \tau \models p \mathbin{\&} q$.

Find a counterexample yourself in which the predicate q is sharp.

4.2.5 Consider a state $\sigma \in \mathcal{D}(X)$, a factor p on X, and a predicate q on X which is non-zero on the support of σ. Show that:

$$\sigma \models \tfrac{p}{q} \geq 1 \Longrightarrow (\sigma \models p) \geq (\sigma \models q),$$

where $\frac{p}{q}(x) = \frac{p(x)}{q(x)}$.

4.2.6 Prove that the following items are equivalent, for a state $\omega \in \mathcal{D}(X)$ and event $E \subseteq X$.

1 $\mathit{supp}(\omega) \subseteq E$;

2 $\omega \models \mathbf{1}_E = 1$;
3 $\omega \models p \,\&\, \mathbf{1}_E = \omega \models p$ for each $p \in Obs(X)$.

4.2.7 Let p_1, p_2, p_3 be predicates on the same set.

1 Show that:

$$(p_1^\perp \otimes p_2^\perp)^\perp = (p_1 \otimes p_2) \mathbin{⩒} (p_1 \otimes p_2^\perp) \mathbin{⩒} (p_1^\perp \otimes p_2).$$

2 Show also that:

$$\begin{aligned}(p_1^\perp \otimes p_2^\perp \otimes p_3^\perp)^\perp = {} & (p_1 \otimes p_2 \otimes p_3) \mathbin{⩒} (p_1 \otimes p_2 \otimes p_3^\perp) \\ & \mathbin{⩒} (p_1 \otimes p_2^\perp \otimes p_3) \mathbin{⩒} (p_1 \otimes p_2^\perp \otimes p_3^\perp) \\ & \mathbin{⩒} (p_1^\perp \otimes p_2 \otimes p_3) \mathbin{⩒} (p_1^\perp \otimes p_2 \otimes p_3^\perp) \\ & \mathbin{⩒} (p_1^\perp \otimes p_2^\perp \otimes p_3).\end{aligned}$$

3 Generalise to n.

4.2.8 For predicates p, q on the same set, define *Reichenbach* implication $\supset$ as:

$$p \supset q \coloneqq p^\perp \mathbin{⩒} (p \,\&\, q).$$

1 Check that:

$$p \supset q = (p \,\&\, q^\perp)^\perp$$

from which it easily follows that:

$$p \supset \mathbf{0} = p^\perp \qquad \mathbf{1} \supset q = q \qquad p \supset \mathbf{1} = \mathbf{1} \qquad \mathbf{0} \supset q = \mathbf{1}.$$

2 Check also that:

$$p^\perp \leq p \supset q \qquad q \leq p \supset q.$$

3 Show that:

$$p_1 \supset (p_2 \supset q) = (p_1 \,\&\, p_2) \supset q.$$

4 For subsets E, D of the same set,

$$\mathbf{1}_E \supset \mathbf{1}_D = \mathbf{1}_{\neg(E \cap \neg D)} = \mathbf{1}_{\neg E \cup D}.$$

The subset $\neg(E \cap \neg D) = \neg E \cup D$ is the standard interpretation of 'E implies D' in Boolean logic (of subsets).

4.2.9 Let p be a predicate on a set X. Prove that the following statements are equivalent.

1 p is sharp;
2 $p \,\&\, p = p$;
3 $p \,\&\, p^\perp = \mathbf{0}$;

4 $q \leq p$ and $q \leq p^{\perp}$ implies $q = \mathbf{0}$, for each $q \in \mathit{Pred}(X)$.

4.2.10 Show that an n-test $p_0, \ldots, p_{n-1}$ on a set X, see Definition 4.2.2, can be identified with a channel $c \colon X \nrightarrow n$, with $p_i = c \mathbin{\overset{!}{\ll}} \mathbf{1}_i$, for $i \in n$.

4.2.11 Let p and q be two arbitrary predicates. Prove that the following use of the partial sum operation $\owedge$ for predicates is justified, that is, yields a well-defined new predicate.

$$(p \otimes q) \owedge (p^{\perp} \otimes q^{\perp}).$$

4.2.12 For a random variable (ω, p), show that the validity of the observable $p - (\omega \models p) \cdot \mathbf{1}$ is zero, *i.e.*,

$$\omega \models (p - (\omega \models p) \cdot \mathbf{1}) \;=\; 0.$$

Observables of this form are used later on to define (co)variance in Section 5.1.

4.2.13 For a multiset $\varphi \in \mathcal{M}(X)$ on a set X and a factor $p \in \mathit{Fact}(X)$ on the same set, define the multiset $\varphi \bullet p \in \mathcal{M}(X)$ as $(\varphi \bullet p)(x) = \varphi(x) \cdot p(x)$. Show that this gives a monoid action:

$$\mathcal{M}(X) \times \mathit{Fact}(X) \xrightarrow{\;\bullet\;} \mathcal{M}(X)$$

wrt. the multiplicative monoid structure $(\mathbf{1}, \&)$ on factors.

4.2.14 Let E be an arbitrary effect algebra. Prove, from the axioms of an effect algebra, that for elements $e, e', d, d', f \in E$ the following properties hold.

1 Orthosupplement is an involution: $e^{\perp\perp} = e$;
2 Cancellation holds of the form: $e \owedge d = e \owedge d'$ implies $d = d'$;
3 Zero-sum freeness holds: $e \owedge d = 0$ implies $e = d = 0$;
4 Define an order $\leq$ on E via: $e \leq d$ iff $e \owedge f = d$ for some $f \in E$. This is a partial order with 1 as top and 0 as bottom element;
5 $e \leq d$ implies $d^{\perp} \leq e^{\perp}$;
6 $e \owedge d$ is defined iff $e \perp d$ iff $e \leq d^{\perp}$ iff $d \leq e^{\perp}$;
7 $e \leq d$ and $d \perp f$ implies $e \perp f$ and $e \owedge f \leq d \owedge f$;
8 if $e \leq e'$ and $d \leq d'$ and $e' \owedge d'$ is defined, then also $e \owedge d$ is defined.

4.2.15 In an effect algebra E, as above, define for elements $e, d \in E$,

$$\begin{array}{lll} e \mathbin{\owedge\!\!\!\ominus} d \;\coloneqq\; (e^{\perp} \owedge d^{\perp})^{\perp} & & \text{if } e^{\perp} \perp d^{\perp} \\ e \ominus d \;\coloneqq\; (e^{\perp} \owedge d)^{\perp} \;=\; e \mathbin{\owedge\!\!\!\ominus} d^{\perp} & & \text{if } e \geq d. \end{array}$$

Show that:

1 $(E, \owedge\!\!\!\ominus, 1)$ is a partial commutative monoid;

2 $e \leq d$ iff $e = d \owedge f$ for some f;
3 $e \ominus 0 = e$ and $1 \ominus e = e^\perp$ and $e \ominus e = 0$;
4 $e \ovee d = f$ iff $e = f \ominus d$; in particular, $(f \ominus d) \ovee d = f$;
5 $e \ovee d \leq f$ iff $e \leq f \ominus d$;
6 $e > d$ iff $e \ominus d > 0$;
7 $f \ominus e = f \ominus d$ implies $e = d$;
8 $e \leq d$ implies $d \ominus e \leq d$ and $d \ominus (d \ominus e) = e$;
9 the function $e \ovee (-)$ preserves all joins $\bigvee_i d_i$ that exist in E: if $e \perp d_i$ for each i, then $e \perp \bigvee_i d_i$ and $e \ovee (\bigvee_i d_i) = \bigvee_i (e \ovee d_i)$;
10 Similarly, $e \owedge (-)$ preserves meets.
11 A homomorphism $f \colon E \to D$ of effect algebras E, D preserves orthosupplement: $f(e^\perp) = f(e)^\perp$, and thus also $f(0) = 0$, $f(e \owedge d) = f(e) \owedge f(d)$ and $f(e \ominus d) = f(e) \ominus f(d)$.

4.2.16 Let E be an effect module. The aim of this exercise is to show that subconvex sums exist in E, see [141, Lemma 2.1]. This means that for arbitrary elements $e_1, \ldots, e_n \in E$ and scalars $r_1, \ldots, r_n \in [0, 1]$ with $\sum_i r_i \leq 1$ the sum $r_1 \cdot e_1 \ovee \cdots \ovee r_n \cdot e_n$ exists in E

1 Use induction on $n \geq 1$, and check the base step.
2 For the induction step let $e_1, \ldots, e_{n+1}$ and $r_1, \ldots, r_{n+1} \in [0, 1]$ with $\sum_{i \leq n+1} r_i \leq 1$ be given. By induction hypothesis the sum $\bigovee_{i \leq n} r_i \cdot e_i$ exists. Use Exercise 4.2.14 to check that the following chain of inequalities holds and suffices to prove that the sum $\bigovee_{i \leq n+1} r_i \cdot e_i$ exists too.

$$\begin{aligned} r_{n+1} \cdot e_{n+1} \;\leq\; r_{n+1} \cdot 1 \;&\leq \Big((\textstyle\sum_{i \leq n} r_i) \cdot 1 \Big)^\perp \\ &= \Big(\textstyle\bigovee_{i \leq n} r_i \cdot 1 \Big)^\perp \;\leq\; \Big(\textstyle\bigovee_{i \leq n} r_i \cdot e_i \Big)^\perp. \end{aligned}$$

4.3 Transformation of observables

One of the basic operations that we have seen so far is state transformation $\gg\!=$. It is used to transform a state ω on the domain X of a channel $c \colon X \to\!\!\!\!\bullet\; Y$ into a state $c \gg\!= \omega$ on the codomain Y of the channel. This section introduces transformation of observables $=\!\ll$. It works in the opposite direction of a channel: an observable q on the codomain Y of the channel is transformed into an observable $c =\!\ll q$ on the domain X of the channel. Thus, state transformation works forwardly, in the direction of the channel, whereas observable transformation $=\!\ll$ works backwardly, against the direction of the channel. This operation $=\!\ll$ is

often applied only to predicates — and then called predicate transformation — but here we apply it more generally to observables.

This section introduces observable transformation $=\!\!\ll$ and lists its key mathematical properties. In the next chapter it will be used for probabilistic reasoning, especially in combination with updating.

Definition 4.3.1. Let $c\colon X \nrightarrow Y$ be a channel. An observable $q \in \mathcal{O}bs(Y)$ is transformed into $c =\!\!\ll q \in \mathcal{O}bs(X)$ via the definition:

$$(c =\!\!\ll q)(x) \;:=\; c(x) \models q \;=\; \sum_{y \in Y} c(x)(y) \cdot q(y).$$

When q is a point predicate $\mathbf{1}_y$, for an element $y \in Y$, we get:

$$(c =\!\!\ll \mathbf{1}_y)(x) \;=\; c(x)(y) \qquad \text{so that} \qquad c =\!\!\ll \mathbf{1}_y \;=\; c(-)(y) : X \to [0,1].$$

The resulting function $Y \to \mathcal{P}red(X)$, given by $y \mapsto c =\!\!\ll \mathbf{1}_y$, is sometimes called the *likelihood* function.

There is a whole series of basic facts about $=\!\!\ll$.

Lemma 4.3.2. *1 The operation $c =\!\!\ll (-)\colon \mathcal{O}bs(Y) \to \mathcal{O}bs(X)$ of transforming observables along a channel $c\colon X \nrightarrow Y$ restricts, first to factors $c =\!\!\ll (-)\colon \mathcal{F}act(Y) \to \mathcal{F}act(X)$, and then to fuzzy predicates $c =\!\!\ll (-)\colon \mathcal{P}red(Y) \to \mathcal{P}red(X)$, but not to sharp predicates.*

2 Observable transformation $c =\!\!\ll (-)$ is linear: it preserves sums $(\mathbf{0}, +)$ of observables and scalar multiplication of observables.

3 Observable transformation preserves truth $\mathbf{1}$, but not conjunction $\&$.

4 Observable transformation preserves the (pointwise) order on observables: $q_1 \leq q_2$ implies $(c =\!\!\ll q_1) \leq (c =\!\!\ll q_2)$.

5 Transformation along the unit channel is the identity: $\mathrm{unit} =\!\!\ll q = q$.

6 Transformation along a composite channel is successive transformation: $(d \circ c) =\!\!\ll q = c =\!\!\ll (d =\!\!\ll q)$.

7 Transformation along a trivial, deterministic channel f is pre-composition: $f =\!\!\ll q = q \circ f$.

Proof. 1 If $q \in \mathcal{F}act(Y)$ then $q(y) \geq 0$ for all y. But then also $(c =\!\!\ll q)(x) = \sum_y c(x)(y) \cdot q(y) \geq 0$, so that $c =\!\!\ll q \in \mathcal{F}act(X)$. If in addition $q \in \mathcal{P}red(Y)$, so that $q(y) \leq 1$ for all y, then also $(c =\!\!\ll q)(x) = \sum_y c(x)(y) \cdot q(y) \leq \sum_y c(x)(y) = 1$, since $c(x) \in \mathcal{D}(Y)$, so that $c =\!\!\ll q \in \mathcal{P}red(X)$.

The fact that a transformation $c =\!\!\ll p$ of a sharp predicate p need not be sharp is demonstrated in Exercise 4.3.2.

2 Easy.

3 $(c \mathrel{=\!\!\ll} \mathbf{1})(x) = \sum_y c(x)(y) \cdot \mathbf{1}(y) = \sum_y c(x)(y) \cdot 1 = 1$. The fact that & is not preserved follows from Exercise 4.3.1.

4 Easy.

5 Recall that $unit(x) = 1|x\rangle$, so that $(unit \mathrel{=\!\!\ll} q)(x) = \sum_y unit(x)(y) \cdot q(y) = q(x)$.

6 For $c \colon X \nrightarrow Y$ and $d \colon Y \nrightarrow Z$ and $q \in Obs(Z)$ we have:

$$\begin{aligned}((d \circ c) \mathrel{=\!\!\ll} q)(x) &= \sum_{z \in Z} (d \circ c)(x)(z) \cdot q(z) \\ &= \sum_{z \in Z} \left(\sum_{y \in Y} c(x)(y) \cdot d(y)(z) \right) \cdot q(z) \\ &= \sum_{y \in Y} c(x)(y) \cdot \left(\sum_{z \in Z} d(y)(z) \cdot q(z) \right) \\ &= \sum_{y \in Y} c(x)(y) \cdot (d \mathrel{=\!\!\ll} q)(y) \\ &= (c \mathrel{=\!\!\ll} (d \mathrel{=\!\!\ll} q))(x).\end{aligned}$$

7 For a function $f \colon X \to Y$,

$$(\langle f \rangle \mathrel{=\!\!\ll} q)(x) = \sum_{y \in Y} unit(f(x))(y) \cdot q(y) = q(f(x)) = (q \circ f)(x). \qquad \blacksquare$$

There is the following fundamental relationship between transformations $\gg\!=$, $=\!\!\ll$ and validity $\models$.

Proposition 4.3.3. *Let $c \colon X \nrightarrow Y$ be a channel with a state $\omega \in \mathcal{D}(X)$ on its domain and an observable $q \in Obs(Y)$ on its codomain. Then:*

$$c \gg\!= \omega \models q \;=\; \omega \models c \mathrel{=\!\!\ll} q. \tag{4.9}$$

Proof. The result follows simply by unpacking the relevant definitions:

$$\begin{aligned}c \gg\!= \omega \models q \;=\; \sum_{y \in Y} (c \gg\!= \omega)(y) \cdot q(y) &= \sum_{y \in Y} \left(\sum_{x \in X} c(x)(y) \cdot \omega(x) \right) \cdot q(y) \\ &= \sum_{x \in X} \omega(x) \cdot \left(\sum_{y \in Y} c(x)(y) \cdot q(y) \right) \\ &= \sum_{x \in X} \omega(x) \cdot (c \mathrel{=\!\!\ll} q)(x) \\ &= \omega \models c \mathrel{=\!\!\ll} q. \qquad \blacksquare\end{aligned}$$

We have already seen several instances of this basic result.

- Earlier we mentioned that marginalisation (of states) and weakening (of observables) are dual to each other, see Equation (4.8). We can now see this as

an instance of (4.9), using a projection $\pi_i \colon X_1 \times \cdots \times X_n \to X_i$ as (trivial) channel, in:

$$\pi_i \gg\!= \omega \models p \;=\; \omega \models \pi_i \mathrel{=\!\!\ll} p$$

On the left-hand side the i-th marginal $\pi_i \gg\!= \omega \in \mathcal{D}(X_i)$ of state $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n)$ is used, whereas on the right-hand side the observable $p \in \mathcal{O}bs(X_i)$ is weakenend to $\pi_i \mathrel{=\!\!\ll} p$ on the product set $X_1 \times \cdots \times X_n$.

- The first equation in Exercise 4.1.4 is also an instance of (4.9), namely for a trivial channel $f \colon X \mathrel{-\!\!\circ} Y$ given by a function $f \colon X \to Y$, as in:

$$\begin{aligned} f \gg\!= \omega \models p &\overset{(4.9)}{=} \omega \models f \mathrel{=\!\!\ll} p \\ &= \omega \models q \circ f \qquad \text{by Lemma 4.3.2 (7).} \end{aligned}$$

Remark 4.3.4. In a programming context, where a channel $c \colon X \mathrel{-\!\!\circ} Y$ is seen as a program taking inputs from X to outputs in Y, one may call $c \mathrel{=\!\!\ll} q$ the *weakest precondition* of q, commonly written as $wp(c, q)$, see *e.g.* [43, 106, 120]. We briefly explain this view.

A precondition of q, w.r.t. channel $c \colon X \mathrel{-\!\!\circ} Y$, may be defined as an observable p on the channel's domain X for which:

$$\omega \models p \;\leq\; c \gg\!= \omega \models q, \qquad \text{for all states } \omega.$$

Proposition 4.3.3 tells that $c \mathrel{=\!\!\ll} q$ is then a precondition of q. It is also the weakest, since if p is a precondition of q, as described above, then in particular:

$$\begin{aligned} p(x) &= \mathit{unit}(x) \models p \\ &\leq c \gg\!= \mathit{unit}(x) \models q \;=\; \mathit{unit}(x) \models c \mathrel{=\!\!\ll} q \;=\; (c \mathrel{=\!\!\ll} q)(x). \end{aligned}$$

As a result, $p \leq c \mathrel{=\!\!\ll} q$.

Lemma 4.3.2 (6) expresses a familiar compositionality property in the theory of weakest preconditions:

$$wp(d \circ c, q) \;=\; (d \circ c) \mathrel{=\!\!\ll} q \;=\; c \mathrel{=\!\!\ll} (d \mathrel{=\!\!\ll} q) \;=\; wp(c, wp(d, q)).$$

We close this section with two topics that dig deeper into the nature of transformations. First, we relate transformation of states and observables in terms of matrix operations. Then we look closer at the categorical aspects of transformation of observables.

Remark 4.3.5. Let c be a channel with finite sets as domain and codomain. For convenience we write these as $n = \{0, \ldots, n-1\}$ and $m = \{0, \ldots, m-1\}$, so that the channel c has type $n \mathrel{-\!\!\circ} m$. For each $i \in n$ we have that $c(i) \in \mathcal{D}(m)$

is given by an m-tuple of numbers in $[0,1]$ that add up to one. Thus we can associate an $m \times n$ matrix M_c with the channel c, namely:

$$M_c = \begin{pmatrix} c(1)(1) & \cdots & c(n-1)(1) \\ \vdots & & \vdots \\ c(1)(m-1) & \cdots & c(n-1)(m-1) \end{pmatrix}.$$

By construction, the columns of this matrix add up to one. Such matrices are often called *stochastic*.

A state $\omega \in \mathcal{D}(n)$ may be identified with a column vector M_ω of length n, as on the left below. It is then easy to see that the matrix $M_{c \gg= \omega}$ of the transformed state, is obtained by *matrix-column multiplication*, as on the right:

$$M_\omega = \begin{pmatrix} \omega(0) \\ \vdots \\ \omega(n-1) \end{pmatrix} \qquad \text{so that} \qquad M_{c \gg= \omega} = M_c \cdot M_\omega.$$

Indeed,

$$(c \gg= \omega)(j) = \sum_i c(i)(j) \cdot \omega(i) = \sum_i (M_c)_{ji} \cdot (M_\omega)_i = (M_c \cdot M_\omega)_j.$$

An observable $q \colon m \to \mathbb{R}$ on m can be identified with a row vector $M_q = \begin{pmatrix} q(0) & \cdots & q(m-1) \end{pmatrix}$. Transformation $c =\!\!\ll q$ then corresponds to *row-matrix multiplication*:

$$M_{c =\!\!\ll q} = M_q \cdot M_c.$$

Again, this is checked easily:

$$(c =\!\!\ll q)(i) = \sum_j q(j) \cdot c(i)(j) = \sum_j (M_q)_j \cdot (M_c)_{j,i} = (M_q \cdot M_c)_i.$$

We close this section by making the functoriality of observable transformation $=\!\!\ll$ explicit, in the style of Proposition 4.2.10. The latter deals with functions, but we now consider functoriality wrt. channels, using the category $\mathbf{Chan} = \mathbf{Chan}(\mathcal{D})$ of probabilistic channels. Notice that wrt. Proposition 4.2.10 the case of sharp predicates is omitted, simply because sharp predicates are not closed under predicate transformation (see Exercise 4.3.2). Also, conjunction & is not preserved under transformation, see Exercise 4.3.1 below.

Proposition 4.3.6. *Taking particular observables on a set is functorial, namely, via functors:*

1 *Obs* $\colon \mathbf{Chan} \to \mathbf{Vect}^{\mathrm{op}}$;
2 *Fact* $\colon \mathbf{Chan} \to \mathbf{Cone}^{\mathrm{op}}$;
3 *Pred* $\colon \mathbf{Chan} \to \mathbf{EMod}^{\mathrm{op}}$;

On a channel $c\colon X \nrightarrow Y$, all these functors are given by transformation $c \mathrel{=\!\!\ll} (-)$, acting in the opposite direction.

Proof. Essentially, all the ingredients are already in Lemma 4.3.2: transformation restricts appropriately (item (1)), transformation preserves identies (item (5)) and composition (item (6)), and the relevant structure (items (2) and (3)). ❑

Exercises

4.3.1 Consider the channel $f\colon \{a, b, c\} \nrightarrow \{u, v\}$ from Example 1.8.2, given by:

$$f(a) = \tfrac{1}{2}|u\rangle + \tfrac{1}{2}|v\rangle \qquad f(b) = 1|u\rangle \qquad f(c) = \tfrac{3}{4}|u\rangle + \tfrac{1}{4}|v\rangle.$$

Take as predicates $p, q\colon \{u, v\} \to [0, 1]$,

$$p(u) = \tfrac{1}{2} \qquad p(v) = \tfrac{2}{3} \qquad q(u) = \tfrac{1}{4} \qquad q(v) = \tfrac{1}{6}.$$

Alternatively, in the normal-form notation of Lemma 4.2.3 (2),

$$p = \tfrac{1}{2} \cdot \mathbf{1}_u + \tfrac{2}{3} \cdot \mathbf{1}_v \qquad\qquad q = \tfrac{1}{4} \cdot \mathbf{1}_u + \tfrac{1}{6} \cdot \mathbf{1}_v.$$

Compute:

- $f \mathrel{=\!\!\ll} p$
- $f \mathrel{=\!\!\ll} q$
- $f \mathrel{=\!\!\ll} (p \mathbin{\mathbb{V}} q)$
- $(f \mathrel{=\!\!\ll} p) \mathbin{\mathbb{V}} (f \mathrel{=\!\!\ll} q)$
- $f \mathrel{=\!\!\ll} (p \mathbin{\&} q)$
- $(f \mathrel{=\!\!\ll} p) \mathbin{\&} (f \mathrel{=\!\!\ll} q)$

This will show that predicate transformation $\mathrel{=\!\!\ll}$ does not preserve conjunction $\&$.

4.3.2 1 Still in the context of the previous exercise, consider the sharp (point) predicate $\mathbf{1}_u$ on $\{u, v\}$. Show that the transformed predicate $f \mathrel{=\!\!\ll} \mathbf{1}_u$ on $\{a, b, c\}$ is not sharp. This proves that sharp predicates are not closed under predicate transformation. This is a good motivation for working with fuzzy predicates instead of with sharp predicates only.

2 Let $h\colon X \to Y$ be a function, considered as deterministic channel, and let $V \subseteq Y$ be an arbitrary subset (event). Check that:

$$h \gg\!= \mathbf{1}_V = \mathbf{1}_{h^{-1}(V)}.$$

Conclude that predicate transformation along deterministic channels does preserve sharpness.

4.3.3 Let $h\colon X \to Y$ be an ordinary function. Recall from Lemma 4.3.2 (7) that $h \mathrel{=\!\!\ll} q = q \circ h$, when h is considered as a deterministic channel. Show that transformation along such deterministic channels does preserve conjunctions:

$$h \mathrel{=\!\!\ll} (q_1 \mathbin{\&} q_2) = (h \mathrel{=\!\!\ll} q_1) \mathbin{\&} (h \mathrel{=\!\!\ll} q_2),$$

in contrast to the findings in Exercise 4.3.1 for arbitrary channels.

Conclude that weakening preserves conjunction: $\mathbf{1} \otimes (p \mathbin{\&} q) = (\mathbf{1} \otimes p) \mathbin{\&} (\mathbf{1} \otimes q)$.

4.3.4 Let predicate $q_i \in \mathcal{P}red(Y)$ form a test, see Definition 4.2.2, and let $c\colon X \mathrel{-\!\!\circ} Y$ be channel. Check that the transformed predicates $c \mathrel{=\!\!\ll} q_i$ form a test on X.

4.3.5 1 Check that a predicate $p\colon X \to [0,1]$ can be identified with a channel $\widehat{p}\colon X \mathrel{-\!\!\circ} 2$, see also Exercise 4.2.10. Describe this $\widehat{p}$ explicitly in terms of p.

2 Define a channel $\mathit{orth}\colon 2 \mathrel{-\!\!\circ} 2$ such that $\mathit{orth} \circ \widehat{p} = \widehat{p^\perp}$.

3 Define also a channel $\mathit{conj}\colon 2 \times 2 \mathrel{-\!\!\circ} 2$ such that $\mathit{conj} \circ \langle \widehat{p}, \widehat{q} \rangle = \widehat{p \mathbin{\&} q}$.

4 Finally, define also a channel $\mathit{scal}(r)\colon 2 \mathrel{-\!\!\circ} 2$, for $r \in [0,1]$, so that $\mathit{scal}(r) \circ \widehat{p} = \widehat{r \cdot p}$.

4.3.6 Recall that a state $\omega \in \mathcal{D}(X)$ can be identified with a channel $1 \mathrel{-\!\!\circ} X$ with a trivial domain, and also that a predicate $p\colon X \to [0,1]$ can be identified with a channel $X \mathrel{-\!\!\circ} 2$, see Exercise 4.3.5. Check that under these identifications validity $\omega \models p$ can be identified with:

- state transformation $p \mathrel{=\!\!\gg} \omega$;
- predicate transformation $\omega \mathrel{=\!\!\ll} p$;
- channel composition $p \circ \omega$.

4.3.7 This exercises shows how parallel conjunction $\otimes$ and sequential conjunction $\mathbin{\&}$ are inter-definable via transformation $\mathrel{=\!\!\ll}$, using projection channels π_i and copy channels Δ, that is, using weakening and contraction.

1 Let observables p_1 on X_1 and p_2 on X_2 be given. Show that on $X_1 \times X_2$,

$$p_1 \otimes p_2 = (\pi_1 \mathrel{=\!\!\ll} p_1) \mathbin{\&} (\pi_2 \mathrel{=\!\!\ll} p_2) = (p_1 \otimes \mathbf{1}) \mathbin{\&} (\mathbf{1} \otimes p_2).$$

The last equation occurred already in Lemma 4.2.9 (2).

2 Let q_1, q_2 be observables on the same set Y. Prove that on Y,

$$q_1 \mathbin{\&} q_2 = \Delta \mathrel{=\!\!\ll} (q_1 \otimes q_2).$$

4.3.8 Prove that:

$$\langle c, d\rangle \mathrel{=\!\!\ll} (p \otimes q) \;=\; (c \mathrel{=\!\!\ll} p) \mathbin{\&} (d \mathrel{=\!\!\ll} q)$$
$$(e \otimes f) \mathrel{=\!\!\ll} (p \otimes q) \;=\; (e \mathrel{=\!\!\ll} p) \otimes (f \mathrel{=\!\!\ll} q).$$

4.3.9 Let $c\colon X \multimap Y$ be a channel, with observables p on Z and q on $Y \times Z$. Check that:

$$(c \otimes id) \mathrel{=\!\!\ll} ((\mathbf{1} \otimes p) \mathbin{\&} q) \;=\; (\mathbf{1} \otimes p) \mathbin{\&} ((c \otimes id) \mathrel{=\!\!\ll} q).$$

(Recall that & is not preserved by $=\!\!\ll$, see Exercise 4.3.1.)

4.3.10 We take a closer look at the functor

$$\mathbf{Chan} \xrightarrow{\mathit{Pred}} \mathbf{EMod}^{\mathrm{op}}$$

from Proposition 4.2.10. It sends a map $c\colon X \multimap Y$ in **Chan** as to $\mathit{Pred}(c)\colon \mathit{Pred}(Y) \to \mathit{Pred}(X)$ via:

$$\mathit{Pred}(c)(q) \;\coloneqq\; c \mathrel{=\!\!\ll} q.$$

1 Check in detail that $\mathit{Pred}(c)$ is a morphism of effect modules, see Definition 4.2.1 (3).
2 Show that the functor *Pred* is *faithful*, in the sense that $\mathit{Pred}(c) = \mathit{Pred}(c')$ implies $c = c'$, for channels $c, c'\colon X \multimap Y$.
3 Let Y be a finite set. Show that for each map $h\colon \mathit{Pred}(Y) \to \mathit{Pred}(X)$ in the category **EMod** there is a unique channel $c\colon X \multimap Y$ with $\mathit{Pred}(c) = h$.
Hint: Write a predicate p as finite sum $\bigotimes_y p(y) \cdot \mathbf{1}_y$, *i.e.* as the normal form of Lemma 4.2.3 (2), and use the relevant preservation property.
One says that the functor *Pred* is *full and faithful* when restricted to the (sub)category with finite sets as objects. In the context of programming (logics) this property is called *healthiness*, see [42, 43, 120], or [63] for an abstract account.

4.4 Validity and drawing

In Chapter 3 we have studied various distributions associated with drawing coloured balls from an urn, such as the multinomial, hypergeometric and Pólya distributions. In this section we look at validity with respect to these draw distributions, in particular at their means.

In Example 4.1.4 (4) we have seen the mean of a binomial distribution. More generally, a multinomial $mn[K](\omega)$ is a distribution on the set $\mathcal{N}[K](X)$ of

natural multisets of size K, and not on (real) numbers. Hence the requirement for a mean, see Definition 4.1.3, does not apply: the space is not a subset of the reals. Still, we have described Proposition 3.3.7 as a generalised mean for multinomials.

Alternatively, we can describe a pointwise mean for multinomials via point-evaluation observables. For a set X with an element $y \in X$ we can define an observable:

$$\mathcal{M}(X) \xrightarrow{\pi_y} \mathbb{R} \qquad \text{by} \qquad \pi_y(\varphi) \coloneqq \varphi(y). \tag{4.10}$$

Then we can compute the validity of this observable as:

$$\begin{aligned} mn[K](\omega) \models \pi_y &\overset{(4.2)}{=} \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \pi_y(\varphi) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \\ &= K \cdot \omega(y), \qquad \text{by Lemma 3.3.4.} \end{aligned} \tag{4.11}$$

In a similary way one gets from Lemma 3.4.4,

$$hg[K](\psi) \models \pi_y \;=\; K \cdot Flrn(\psi)(y) \;=\; pl[K](\psi) \models \pi_y. \tag{4.12}$$

Based on the above equations we define the mean of draw distributions as a multiset of point-validities.

Definition 4.4.1. Let X be a set (of colours) and K be a number. The means of the three draw distributions are defined as multisets over X in:

$$\begin{aligned} mean\Big(mn[K](\omega)\Big) \coloneqq \sum_{y \in X} (mn[K](\omega) \models \pi_y) \,\big|y\big\rangle &= K \cdot \sum_{y \in X} \omega(y) \,\big|y\big\rangle \\ &= K \cdot \omega. \\ mean\Big(hg[K](\psi)\Big) \coloneqq \sum_{y \in X} (hg[K](\psi) \models \pi_y) \,\big|y\big\rangle &= K \cdot \sum_{y \in X} Flrn(\psi)(y) \,\big|y\big\rangle \\ &= K \cdot Flrn(\psi) \\ mean\Big(pl[K](\psi)\Big) \coloneqq \sum_{y \in X} (pl[K](\psi) \models \pi_y) \,\big|y\big\rangle &= K \cdot \sum_{y \in X} Flrn(\psi)(y) \,\big|y\big\rangle \\ &= K \cdot Flrn(\psi). \end{aligned}$$

These formulations coincide with the ones that we have given earlier in Propositions 3.3.7 and 3.4.7.

We recall a historical example, known as the paradox of the Chevalier De Méré, from the 17th century. He argued informally that the following two outcomes have equal probability.

DM1 Throw a dice 4 times; you get at least one 6.
DM2 Throw a pair of dice 24 times; you get at least one double six, *i.e.* $(6, 6)$.

However, when De Méré betted on (*DM2*) he mostly lost. Puzzled, he wrote to Pascal, who showed that the probabilities differ.

Let us write $(-)(6) \geq 1 \colon \mathcal{N}(\mathit{pips}) \to \{0, 1\}$ for the sharp predicate that sends a multiset φ, over the dice space $\mathit{pips} = \{1, 2, 3, 4, 5, 6\}$, to 1 if $\varphi(6) \geq 1$ and to 0 otherwise. It thus tells that the number 6 occurs at least once in the draw φ. We can thus model the first option (*DM1*) as validity:

$$\mathit{mn}[4](\mathit{dice}) \models (-)(6) \geq 1. \tag{*DM1*}$$

The second option (*DM2*) then becomes:

$$\mathit{mn}[24](\mathit{dice} \otimes \mathit{dice}) \models (-)(6, 6) \geq 1, \tag{*DM2*}$$

where $(-)(6, 6) \geq 1 \colon \mathcal{N}(\mathit{dice} \times \mathit{dice}) \to \{0, 1\}$ is the obvious predicate.

Using a computer to calculate the validity (*DM1*) is relatively easy, giving an outcome $\frac{671}{1296}$. It involves summing over $\left(\!\binom{6}{4}\!\right) = 126$ multisets, see Proposition 1.7.4. However, the sum in (*DM1*) is huge, involving $\left(\!\binom{36}{24}\!\right)$ multisets — a number in the order of $2 \cdot 10^{16}$.

The common way to compute (*DM1*) and (*DM2*) is to switch to validity over the distribution *dice*, and using orthosupplement (negations). Getting at least one 6 in 4 throws is the orthosupplement of getting 4 times no 6. This can be represented and computed as:

$$\begin{aligned} \mathit{dice} \otimes \mathit{dice} \otimes \mathit{dice} \otimes \mathit{dice} \models (\mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp)^\perp &= 1 - \left(\mathit{dice} \models \mathbf{1}_6^\perp\right)^4 \\ &= 1 - (\tfrac{5}{6})^4 \\ &= \tfrac{671}{1296} \\ &\approx 0.518. \end{aligned}$$

Similarly one can compute (*DM2*) as:

$$\begin{aligned} (\mathit{dice} \otimes \mathit{dice})^{24} \models \left(\left((\mathbf{1}_6 \otimes \mathbf{1}_6)^\perp\right)^{24}\right)^\perp &= 1 - (\tfrac{35}{36})^{24} \\ &\approx 0.491. \end{aligned}$$

Hence indeed, betting on (*DM2*) is a bad idea.

In this section we look at validities over distributions of draws from an urn, like in (*DM1*) and (*DM2*). In particular, we establish connections between such validities and validities over the urn, as distribution. This involves free extensions of observables to multisets.

We notice that there are two ways to extend an observable $X \to \mathbb{R}$ on a set X to natural multisets over X, since we can choose to use either the additive

structure or the multiplicative structure on $\mathbb{R}$. Both these extensions are based on Proposition 1.4.4.

Definition 4.4.2. Let $p\colon X \to \mathbb{R}$ be an observable on a set X.

1 The *additive extension* $\overline{p}^{+}\colon \mathcal{N}(X) \to \mathbb{R}$ of p on multisets over X is defined as:

$$\overline{p}^{+}(\varphi) = \sum_{x \in supp(\varphi)} \varphi(x) \cdot p(x) = \sum_{x \in X} \varphi(x) \cdot p(x).$$

2 The *multiplicative extension* $\overline{p}^{\bullet}\colon \mathcal{N}(X) \to \mathbb{R}$ of p is:

$$\overline{p}^{\bullet}(\varphi) = \prod_{x \in supp(\varphi)} p(x)^{\varphi(x)} = \prod_{x \in X} p(x)^{\varphi(x)}.$$

By construction, these extensions are homomorphisms of monoids, so that:

$$\begin{cases} \overline{p}^{+}(\mathbf{0}) = 0 \\ \overline{p}^{\bullet}(\mathbf{0}) = 1. \end{cases} \quad \text{and} \quad \begin{cases} \overline{p}^{+}(\varphi + \psi) = \overline{p}^{+}(\varphi) + \overline{p}^{+}(\psi) \\ \overline{p}^{\bullet}(\varphi + \psi) = \overline{p}^{\bullet}(\varphi) \cdot \overline{p}^{\bullet}(\psi) \end{cases} \tag{4.13}$$

Once extended to multisets, we can investigate the validity of these extensions in distributions obtained from drawing. We concentrate on the multinomial case first where the validity of both the additive and the multiplicative extensions can be formulated in terms of the validity in the underlying state (urn).

Proposition 4.4.3. *Consider a random variable given by an observable* $p\colon X \to \mathbb{R}$ *and a state* $\omega \in \mathcal{D}(X)$*. Fix* $K \in \mathbb{N}$*.*

1 The additive extension $\overline{p}^{+}$ *of* p *forms new random variable, with the multinomial distribution* $mn[K](\omega)$ *on* $\mathcal{N}[K](X)$ *as state. The associated validity is:*

$$mn[K](\omega) \models \overline{p}^{+} = K \cdot (\omega \models p).$$

2 The multiplicative extension gives:

$$mn[K](\omega) \models \overline{p}^{\bullet} = (\omega \models p)^{K}.$$

Proof. 1 We use Lemma 3.3.4 in:

$$\begin{aligned} mn[K](\omega) \models \overline{p}^{+} &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \overline{p}^{+}(\varphi) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \left(\sum_{x \in X} \varphi(x) \cdot p(x) \right) \\ &= \sum_{x \in X} p(x) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x) \\ &= \sum_{x \in X} p(x) \cdot K \cdot \omega(x) \\ &= K \cdot (\omega \models p). \end{aligned}$$

2 By unfolding the multinomial distribution, see (3.9):

$$\begin{aligned} mn[K](\omega) \models \overline{p}^{\bullet} &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \overline{p}^{\bullet}(\varphi) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} \binom{K}{\varphi} \cdot \left(\prod\nolimits_x \omega(x)^{\varphi(x)} \right) \cdot \left(\prod\nolimits_x p(x)^{\varphi(x)} \right) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} \binom{K}{\varphi} \cdot \prod\nolimits_x (\omega(x) \cdot p(x))^{\varphi(x)} \\ &\overset{(1.26)}{=} \left(\sum\nolimits_x \omega(x) \cdot p(x) \right)^{K} \\ &= (\omega \models p)^{K}. \end{aligned}$$

❑

For the hypergeometric and Pólya distributions we have results for the additive extension. They resemble the formulation in the above first item.

Proposition 4.4.4. *Consider an observable p and an urn ψ.*

$$hg[K](\psi) \models \overline{p}^{+} \;=\; K \cdot (Flrn(\psi) \models p) \;=\; pl[K](\psi) \models \overline{p}^{+}.$$

Proof. Both equations are obtained as for Proposition 4.4.3 (1), this time using Lemma 3.4.4. ❑

For the parallel multinomial law $pml \colon \mathcal{M}[K](\mathcal{D}(X)) \to \mathcal{D}(\mathcal{M}[K](X))$ from Section 3.6 there is a similar result, in the multiplicative case.

Proposition 4.4.5. *For an observable $p \colon X \to \mathbb{R}$ and for a multiset of distributions $\sum_i n_i |\omega_i\rangle \in \mathcal{M}[K](\mathcal{D}(X))$,*

$$pml(\textstyle\sum_i n_i |\omega_i\rangle) \models \overline{p}^{\bullet} \;=\; \prod\nolimits_i (\omega_i \models p)^{n_i}.$$

Proof. We use the second formulation (3.25) of *pml* in:

$$\begin{aligned}
pml(\textstyle\sum_i n_i|\omega_i\rangle) \models \overline{p}^\bullet &= \sum_{i,\varphi_i \in \mathcal{N}[n_i](X)} \Big(\textstyle\prod_i\ mn[n_i](\omega_i)(\varphi_i)\Big) \cdot \overline{p}^\bullet(\textstyle\sum_i \varphi_i) \\
&= \sum_{i,\varphi_i \in \mathcal{N}[n_i](X)} \Big(\textstyle\prod_i\ mn[n_i](\omega_i)(\varphi_i)\Big) \cdot \Big(\textstyle\prod_i \overline{p}^\bullet(\varphi_i)\Big) \\
&\overset{(4.13)}{=} \sum_{i,\varphi_i \in \mathcal{N}[n_i](X)} \textstyle\prod_i\ mn[n_i](\omega_i)(\varphi_i) \cdot \overline{p}^\bullet(\varphi_i) \\
&= \textstyle\prod_i \displaystyle\sum_{\varphi_i \in \mathcal{N}[n_i](X)} mn[n_i](\omega_i)(\varphi_i) \cdot \overline{p}^\bullet(\varphi_i) \\
&= \textstyle\prod_i mn[n_i](\omega_i) \models \overline{p}^\bullet \\
&= \textstyle\prod_i (\omega_i \models p)^{n_i}, \qquad \text{by Proposition 4.4.3 (2).} \quad \square
\end{aligned}$$

We conclude this section by taking a fresh look at multinomial distributions. For convenience, we recall the formulation from (3.9), for an distribution (as urn) $\omega \in \mathcal{D}(X)$ and number $K \in \mathbb{N}$:

$$mn[K](\omega) \;=\; \sum_{\varphi \in \mathcal{N}[K](X)} \binom{K}{\varphi} \cdot \textstyle\prod_x \omega(x)^{\varphi(x)} \,\big|\varphi\big\rangle,$$

The probability $\omega(x)$ equals the validity $\omega \models \mathbf{1}_x$ of the point predicate $\mathbf{1}_x$, for the element $x \in X$ with multiplicity $\varphi(x)$ in the mulitset (draw) φ. We ask ourselves the question: can we replace these point predicates with arbitrary predicates? Does such a generalisation make sense?

As we shall see in Chapter **??** on probabilistic learning, this makes perfectly good sense. It involves a generalisation of multisets over points, as elements of a set X, to multisets over predicates on X. Such predicates can express uncertainties over the data (elements) from which we wish to learn.

In the current setting we describe only how to adapt multinomial distributions to predicates. This works for tests, which are finite sets of predicates adding up to the truth predicate $\mathbf{1}$, see Definition 4.2.2. It turns out that there are two forms of "logical" multinomial for predicates.

Definition 4.4.6. Fix a state $\omega \in \mathcal{D}(X)$ and a number $K \in \mathbb{N}$. Let $T \subseteq Pred(X)$ be a test.

1 Define the *external logical* multinomial as:

$$mn_E[K](\omega, T) \;=\; \sum_{\varphi \in \mathcal{N}[K](T)} \binom{K}{\varphi} \cdot \prod_{p \in T} (\omega \models p)^{\varphi(p)} \,\big|\varphi\big\rangle,$$

2 There is also the *internal logical* multinomial:

$$mn_I[K](\omega, T) \;=\; \sum_{\varphi \in \mathcal{N}[K](P)} \multinom{\varphi} \cdot \Big(\omega \models \mathop{\&}_{p \in T} p^{\varphi(p)} \Big) \,|\varphi\rangle,$$

In the first, external case the product $\prod$ is outside the validity $\models$, whereas in the second, internal case the conjunction (product) & is inside the validity $\models$. This explains the names.

It is easy to see how the internal logical formulation generalises the standard multinomial distribution. For a distribution ω with support $X = \mathit{supp}(\omega)$, say of the form $X = \{x_1, \ldots, x_n\}$, the point predicates $\mathbf{1}_{x_1}, \ldots, \mathbf{1}_{x_n}$ form a test on X, since $\mathbf{1}_{x_1} \otimes \cdots \otimes \mathbf{1}_{x_n} = \mathbf{1}$. A multiset over these predicates $\mathbf{1}_{x_1}, \ldots, \mathbf{1}_{x_n}$ can easily be identified with a multiset over the points $x_1, \ldots, x_n$.

The external logical formulation only makes sense for proper fuzzy predicates. In the sharp case, say with a test $\mathbf{1}_U, \mathbf{1}_{\neg U}$ things trivialise, as in:

$$(\mathbf{1}_U)^n \mathbin{\&} (\mathbf{1}_{\neg U})^m \;=\; \mathbf{1}_U \mathbin{\&} \mathbf{1}_{\neg U} \;=\; \mathbf{0}.$$

It is probably for this reason that the internal version is not so common. But it is quite natural in Bayesian learning, see Chapter **??**.

We need to check that the above definitions yield actual distributions, with probabilities adding up to one. In both cases this follows from the Multinomial Theorem (1.27). Externally:

$$\begin{aligned} \sum_{\varphi \in \mathcal{N}[K](P)} \multinom{\varphi} \cdot \prod_{p \in P} (\omega \models p)^{\varphi(p)} &\overset{(1.27)}{=} \left(\sum_{p \in P} \omega \models p \right)^K \;=\; \left(\omega \models \sum_{p \in P} p \right)^K \\ &= \big(\omega \models \mathbf{1}\big)^K \;=\; 1^K \;=\; 1. \end{aligned}$$

In the internal case we get:

$$\begin{aligned} \sum_{\varphi \in \mathcal{N}[K](P)} \multinom{\varphi} \cdot \Big(\omega \models \mathop{\&}_{p \in P} p^{\varphi(p)} \Big) \;&=\; \omega \models \sum_{\varphi \in \mathcal{N}[K](P)} \multinom{\varphi} \cdot \Big(\mathop{\&}_{p \in P} p^{\varphi(p)} \Big) \\ &= \sum_{x \in X} \omega(x) \cdot \sum_{\varphi \in \mathcal{N}[K](P)} \multinom{\varphi} \cdot \prod_{p \in P} p(x)^{\varphi(p)} \\ &\overset{(1.27)}{=} \sum_{x \in X} \omega(x) \cdot \left(\sum_{p \in P} p(x) \right)^K \;=\; \sum_{x \in X} \omega(x) \cdot 1^K \;=\; 1. \end{aligned}$$

Exercises

4.4.1 Show that the additive/multiplicative extension preserves the additive/multiplicative structure of observables:

$$\overline{p + q}^{+} \;=\; \overline{p}^{+} + \overline{q}^{+} \qquad\qquad \overline{\mathbf{0}}^{+} \;=\; \mathbf{0}$$

and:

$$\overline{p \mathbin{\&} q}^{\bullet} = \overline{p}^{\bullet} \mathbin{\&} \overline{q}^{\bullet} \qquad \overline{\mathbf{1}}^{\bullet} = \mathbf{1}$$

4.4.2 For a random variable $p\colon X \to \mathbb{R}$ with state $\omega \in \mathcal{D}(X)$ consider the validity $mn[-](\omega) \models \overline{p}^{+}$ as an observable $\mathbb{N} \to \mathbb{R}$. Show that for a distribution $\sigma \in \mathcal{D}(\mathbb{N})$ one has:

$$\sigma \models \Big(mn[-](\omega) \models \overline{p}\Big) = \mathit{mean}(\sigma) \cdot (\omega \models p).$$

4.4.3 For a random variable (ω, p) write $\sum(\omega, p)\colon \mathbb{N} \to \mathbb{R}$ for the summation observable defined by:

$$\textstyle\sum(\omega, p)(n) \coloneqq \omega \models p + \cdots + p \quad (n \text{ times}).$$

Show that for a distribution $\sigma \in \mathcal{D}(\mathbb{N})$ one has:

$$\textstyle\sigma \models \sum(\omega, p) = \mathit{mean}(\sigma) \cdot (\omega \models p).$$

4.4.4 The following is often used as illustration of Wald's identity. Roll a dice, and let $n \in \mathit{pips} = \{1, \ldots, 6\}$ be the number that comes up; then roll the dice n more times and record the sum of the resulting pips. What is the expected sum?

1 Use Exercise 4.4.3 to show that the expected number if $\frac{49}{4}$.

2 Obtain this same outcome via Proposition 4.4.3 (1).

4.4.5 Consider the set $X = \{a, b\}$ with state $\omega = \frac{1}{3}|a\rangle + \frac{2}{3}|b\rangle$ and 3-test of predicates $T = \{p_1, p_2, p_3\}$ where:

$$p_1 = \tfrac{1}{8} \cdot \mathbf{1}_a + \tfrac{1}{2} \cdot \mathbf{1}_b \qquad p_2 = \tfrac{1}{4} \cdot \mathbf{1}_a + \tfrac{1}{6} \cdot \mathbf{1}_b \qquad p_2 = \tfrac{5}{8} \cdot \mathbf{1}_a + \tfrac{1}{3} \cdot \mathbf{1}_b.$$

1 Show that:

$$\begin{aligned} &mn_E[2](\omega, T) \\ &= \tfrac{9}{64}\big|2|p_1\rangle\big\rangle + \tfrac{7}{48}\big|1|p_1\rangle + 1|p_2\rangle\big\rangle + \tfrac{49}{1296}\big|2|p_2\rangle\big\rangle \\ &\qquad + \tfrac{31}{96}\big|1|p_1\rangle + 1|p_3\rangle\big\rangle + \tfrac{217}{1296}\big|1|p_2\rangle + 1|p_3\rangle\big\rangle + \tfrac{961}{5184}\big|2|p_3\rangle\big\rangle. \end{aligned}$$

2 And similarly that:

$$\begin{aligned} &mn_I[2](\omega, T) \\ &= \tfrac{11}{64}\big|2|p_1\rangle\big\rangle + \tfrac{19}{144}\big|1|p_1\rangle + 1|p_2\rangle\big\rangle + \tfrac{17}{432}\big|2|p_2\rangle\big\rangle \\ &\qquad + \tfrac{79}{288}\big|1|p_1\rangle + 1|p_3\rangle\big\rangle + \tfrac{77}{432}\big|1|p_2\rangle + 1|p_3\rangle\big\rangle + \tfrac{353}{1728}\big|2|p_3\rangle\big\rangle. \end{aligned}$$

4.5 Validity-based distances

This section describes two standard notions of distance between states, and also between predicates. It shows that these distances can both be formulated in terms of validity $\models$, in a dual form. Basic properties of these distances are included. Earlier, in Section 2.7, we have have seen Kullback-Leibler divergence as a measure of difference between states. But divergence does not form a distance function since it is not symmetric, see Remark 4.5.5 for an illustration of the difference.

The distance between two states $\omega_1, \omega_2 \in \mathcal{D}(X)$, on the same set X, can be defined as the join of the distances in $[0, 1]$, between validities:

$$d(\omega_1, \omega_2) \coloneqq \bigvee_{p \in Pred(X)} \big|\omega_1 \models p \;-\; \omega_2 \models p\big|. \tag{4.14}$$

Similarly, for two predicates $p_1, p_2 \in Pred(X)$, one can define:

$$d(p_1, p_2) \coloneqq \bigvee_{\omega \in \mathcal{D}(X)} \big|\omega \models p_1 \;-\; \omega \models p_2\big|. \tag{4.15}$$

Note that the above formulations involve predicates only, not observables in general.

4.5.1 Distance between states

The distance defined in (4.14) is commonly called the *total variation* distance, which is a special case of the Kantorovich distance, see *e.g.* [56, 15, 121, 117]. Its two alternative characterisations below are standard. We refer to [86] for more information about the validity-based approach.

Proposition 4.5.1. *Let X be an arbitrary set, with states $\omega_1, \omega_2 \in \mathcal{D}(X)$. Then:*

$$d(\omega_1, \omega_2) \;=\; \max_{U \subseteq X} \omega_1 \models \mathbf{1}_U \;-\; \omega_2 \models \mathbf{1}_U \;=\; \tfrac{1}{2} \sum_{x \in X} \big|\omega_1(x) \;-\; \omega_2(x)\big|.$$

We write maximum 'max' instead of join $\bigvee$ to express that the supremum is actually reached by a subset (sharp predicate).

Proof. Let $\omega_1, \omega_2 \in \mathcal{D}(X)$ be two discrete probability distributions on the same set X. We will prove the two inequalities labeled (a) and (b) in:

$$\begin{aligned} \tfrac{1}{2} \textstyle\sum_{x \in X} \big|\omega_1(x) \;-\; \omega_2(x)\big| &\overset{(a)}{\leq} \max_{U \subseteq X} \omega_1 \models \mathbf{1}_U - \omega_2 \models \mathbf{1}_U \\ &\leq \bigvee_{p \in Pred(X)} \big|\omega_1 \models p - \omega_2 \models p\big| \\ &\overset{(b)}{\leq} \tfrac{1}{2} \sum_{x \in X} \big|\omega_1(x) \;-\; \omega_2(x)\big|. \end{aligned}$$

This proves Proposition 4.5.1 since the inequality in the middle is trivial.

We start with some preparatory definitions. Let $U \subseteq X$ be an arbitrary subset. We shall write $\omega_i(U) = \sum_{x\in U} \omega_i(x) = (\omega \models \mathbf{1}_U)$. We partition U in three disjoint parts, and take the relevant sums:

$$\begin{cases} U_> = \{x \in U \mid \omega_1(x) > \omega_2(x)\} \\ U_= = \{x \in U \mid \omega_1(x) = \omega_2(x)\} \\ U_< = \{x \in U \mid \omega_1(x) < \omega_2(x)\} \end{cases} \qquad \begin{cases} U\uparrow = \omega_1(U_>) - \omega_2(U_>) \geq 0 \\ U\downarrow = \omega_2(U_<) - \omega_1(U_<) \geq 0. \end{cases}$$

We use this notation in particular for $U = X$. In that case we can use:

$$\begin{aligned} 1 = \omega_1(X) &= \omega_1(X_>) + \omega_1(X_=) + \omega_1(X_<) \\ 1 = \omega_2(X) &= \omega_2(X_>) + \omega_2(X_=) + \omega_2(X_<) \end{aligned}$$

Hence by subtraction we obtain, since $\omega_1(X_=) = \omega_2(X_=)$,

$$0 = (\omega_1(X_>) - \omega_2(X_>)) + (\omega_1(X_<) - \omega_2(X_<))$$

That is,

$$X\uparrow = \omega_1(X_>) - \omega_2(X_>) = \omega_2(X_<) - \omega_1(X_<) = X\downarrow .$$

As a result:

$$\begin{aligned} &\tfrac{1}{2} \textstyle\sum_{x\in X} \big|\, \omega_1(x) - \omega_2(x) \big| \\ &= \tfrac{1}{2}\Big(\textstyle\sum_{x\in X_>} (\, \omega_1(x) - \omega_2(x)) + \sum_{x\in X_<} (\, \omega_2(x) - \omega_1(x))\Big) \\ &= \tfrac{1}{2}\Big((\, \omega_1(X_>) - \omega_2(X_>)) + (\, \omega_2(X_<) - \omega_1(X_<))\Big) \\ &= \tfrac{1}{2}(X\uparrow + X\downarrow) \\ &= X\uparrow \end{aligned} \tag{4.16}$$

We have prepared the ground for proving the above inequalities (a) and (b).

(a) We will see that the above maximum is actually reached for the subset $U = X_>$, first of all because:

$$\begin{aligned} \tfrac{1}{2} \textstyle\sum_{x\in X} \big|\, \omega_1(x) - \omega_2(x) \big| \overset{(4.16)}{=} X\uparrow &= \omega_1(X_>) - \omega_2(X_>) \\ &= \omega_1 \models \mathbf{1}_{X_>} - \omega_2 \models \mathbf{1}_{X_>} \\ &\leq \max_{U\subseteq X} \omega_1 \models \mathbf{1}_U - \omega_2 \models \mathbf{1}_U. \end{aligned}$$

(b) Let $p \in \mathit{Pred}(X)$ be an arbitrary predicate. We have: $(\mathbf{1}_U \;\&\; p)(x) = \mathbf{1}_U(x) \cdot$

$p(x)$, which is $p(x)$ if $x \in U$ and 0 otherwise. Then:

$$\begin{aligned}
&\big|\omega_1 \models p - \omega_2 \models p\big| \\
&= \Big|(\omega_1 \models \mathbf{1}_{X_>} \& p \;+\; \omega_1 \models \mathbf{1}_{X_=} \& p \;+\; \omega_1 \models \mathbf{1}_{X_<} \& p) \\
&\qquad - (\omega_2 \models \mathbf{1}_{X_>} \& p \;+\; \omega_2 \models \mathbf{1}_{X_=} \& p \;+\; \omega_2 \models \mathbf{1}_{X_<} \& p)\Big| \\
&= \Big|(\omega_1 \models \mathbf{1}_{X_>} \& p \;-\; \omega_2 \models \mathbf{1}_{X_>} \& p) \;-\; (\omega_2 \models \mathbf{1}_{X_<} \& p \;-\; \omega_1 \models \mathbf{1}_{X_<} \& p)\Big| \\
&= \begin{cases} (\omega_1 \models \mathbf{1}_{X_>} \& p \;-\; \omega_2 \models \mathbf{1}_{X_>} \& p) \;-\; (\omega_2 \models \mathbf{1}_{X_<} \& p \;-\; \omega_1 \models \mathbf{1}_{X_<} \& p) \\ \qquad \text{if } \omega_1 \models \mathbf{1}_{X_>} \& p - \omega_2 \models \mathbf{1}_{X_>} \& p \overset{(*)}{\geq} \omega_2 \models \mathbf{1}_{X_<} \& p - \omega_1 \models \mathbf{1}_{X_<} \& p \\ (\omega_2 \models \mathbf{1}_{X_<} \& p \;-\; \omega_1 \models \mathbf{1}_{X_<} \& p) \;-\; (\omega_1 \models \mathbf{1}_{X_>} \& p \;-\; \omega_2 \models \mathbf{1}_{X_>} \& p) \\ \qquad \text{otherwise} \end{cases} \\
&\leq \begin{cases} \omega_1 \models \mathbf{1}_{X_>} \& p \;-\; \omega_2 \models \mathbf{1}_{X_>} \& p & \text{if } (*) \\ \omega_2 \models \mathbf{1}_{X_<} \& p \;-\; \omega_1 \models \mathbf{1}_{X_<} \& p & \text{otherwise} \end{cases} \\
&= \begin{cases} \sum_{x \in X_>} (\omega_1(x) - \omega_2(x)) \cdot p(x) & \text{if } (*) \\ \sum_{x \in X_<} (\omega_2(x) - \omega_1(x)) \cdot p(x) & \text{otherwise} \end{cases} \\
&\leq \begin{cases} \sum_{x \in X_>} \omega_1(x) - \omega_2(x) & \text{if } (*) \\ \sum_{x \in X_<} \omega_2(x) - \omega_1(x) & \text{otherwise} \end{cases} \\
&= \begin{cases} X\uparrow & \text{if } (*) \\ X\downarrow \;=\; X\uparrow & \text{otherwise} \end{cases} \\
&= X\uparrow \\
&\overset{(4.16)}{=} \tfrac{1}{2} \sum_{x \in X} \big|\, \omega_1(x) - \omega_2(x) \,\big|.
\end{aligned}$$

This completes the proof. ❑

The sum-formulation in Proposition 4.5.1 is useful in many situations, for instance in order to prove that the above distance function d between states is a metric.

Lemma 4.5.2. *The distance $d(\omega_1, \omega_2)$ between states $\omega_1, \omega_2 \in \mathcal{D}(X)$ in* (4.14) *turns the set of distributions $\mathcal{D}(X)$ into a metric space, with* $[0, 1]$*-valued metric.*

Proof. If $d(\omega_1, \omega_2) = \frac{1}{2} \sum_{x \in X} |\, \omega_1(x) - \omega_2(x) \,| = 0$, then $|\, \omega_1(x) - \omega_2(x) \,| = 0$ for each $x \in X$, so that $\omega_1(x) = \omega_2(x)$, and thus $\omega_1 = \omega_2$. Obviously, $d(\omega_1, \omega_2) = d(\omega_2, \omega_1)$. The triangle inequality holds for d since it holds for the standard

distance on $[0, 1]$.

$$\begin{aligned} d(\omega_1, \omega_3) &= \tfrac{1}{2} \sum_{x \in X} |\, \omega_1(x) - \omega_3(x) \,| \\ &\leq \tfrac{1}{2} \sum_{x \in X} |\, \omega_1(x) - \omega_2(x) \,| + |\, \omega_2(x) - \omega_3(x) | \\ &= \tfrac{1}{2} \sum_{x \in X} |\, \omega_1(x) - \omega_2(x) \,| + \tfrac{1}{2} \sum_{x \in X} |\, \omega_2(x) - \omega_3(x) \,| \\ &= d(\omega_1, \omega_2) + d(\omega_2, \omega_3). \end{aligned}$$ ❑

We use this same sum-formulation for the following result.

Lemma 4.5.3. *State transformation is non-expansive: for a channel $c \colon X \nrightarrow Y$ one has:*

$$d(c \gg\!= \omega_1, c \gg\!= \omega_2) \;\leq\; d(\omega_1, \omega_2).$$

Proof. Since:

$$\begin{aligned} d(c \gg\!= \omega_1, c \gg\!= \omega_2) &= \tfrac{1}{2} \sum_y \big| (c \gg\!= \omega_1)(y) - (c \gg\!= \omega_2(y) \big| \\ &= \tfrac{1}{2} \sum_y \Big| \sum_x \omega_1(x) \cdot c(x)(y) - \sum_x \omega_2(x) \cdot c(x)(y) \Big| \\ &= \tfrac{1}{2} \sum_y \Big| \sum_x c(x)(y) \cdot (\omega_1(x) - \omega_2(x)) \Big| \\ &\leq \tfrac{1}{2} \sum_y \sum_x c(x)(y) \cdot \big| \omega_1(x) - \omega_2(x) \big| \\ &= \tfrac{1}{2} \sum_x \big(\sum_y c(x)(y) \big) \cdot \big| \omega_1(x) - \omega_2(x) \big| \\ &= d(\omega_1, \omega_2). \end{aligned}$$ ❑

For two states $\omega_1, \omega_2 \in \mathcal{D}(X)$ a *coupling* is a joint state $\sigma \in \mathcal{D}(X \times X)$ that marginalises to ω_1 and ω_2, *i.e.* that satisfies $\sigma[1, 0] = \omega_1$ and $\sigma[0, 1] = \omega_2$. Such couplings give an alternative formulation of the distance between states, which is often called the Wasserstein distance, see [164]. These couplings will also be used in Section 7.8 as probabilistic relations. The proof of the next result is standard and is included in order to be complete.

Proposition 4.5.4. *For states $\omega_1, \omega_2 \in \mathcal{D}(X)$,*

$$d(\omega_1, \omega_2) \;=\; \bigwedge \{ \sigma \models Eq^{\perp} \mid \sigma \text{ is a coupling between } \omega_1, \omega_2 \},$$

where $Eq \colon X \times X \to [0, 1]$ is the equality predicate from Definition 4.1.1 (4)*.*

Proof. We use the notation and results from the proof of Proposition 4.5.1. We first prove the inequality ($\leq$). Let $X_> = \{x \in X \mid \omega_1(x) > \omega_2(x)\}$. Then:

$$\begin{aligned} \omega_1(x) \;=\; \textstyle\sum_y \sigma(x, y) \;&=\; \sigma(x, x) + \textstyle\sum_{y \neq x} \sigma(x, y) \\ &\leq\; \omega_2(x) + (\sigma \models (\mathbf{1}_x \otimes \mathbf{1}) \;\&\; Eq^{\perp}). \end{aligned}$$

This means that $\omega_1(x) - \omega_2(x) \leq \sigma \models (\mathbf{1}_x \otimes \mathbf{1}) \;\&\; Eq^{\perp}$ for $x \in X_>$. We similarly

have:

$$\begin{aligned}\omega_2(x) \;=\; \textstyle\sum_x \sigma(y,x) \;&=\; \sigma(x,x) + \textstyle\sum_{y\neq x} \sigma(y,x) \\ &\leq\; \omega_1(x) + (\sigma \models (\mathbf{1} \otimes \mathbf{1}_x) \;\&\; Eq^{\perp}).\end{aligned}$$

Hence $\omega_2(x) - \omega_1(x)$ for $x \notin X_>$. Putting this together gives:

$$\begin{aligned}d(\omega_1,\omega_2) \;&=\; \tfrac{1}{2} \textstyle\sum_{x\in X} \big|\, \omega_1(x) - \omega_3(x) \,\big| \\ &=\; \tfrac{1}{2} \textstyle\sum_{x\in X_>} \omega_1(x) - \omega_2(x) \;+\; \tfrac{1}{2} \textstyle\sum_{x\in \neg X_>} \omega_2(x) - \omega_1(x) \\ &\leq\; \tfrac{1}{2} \textstyle\sum_{x\in X_>} \sigma \models (\mathbf{1}_x \otimes \mathbf{1}) \;\&\; Eq^{\perp} \;+\; \tfrac{1}{2} \textstyle\sum_{x\in \neg X_>} \sigma \models (\mathbf{1} \otimes \mathbf{1}_x) \;\&\; Eq^{\perp} \\ &=\; \tfrac{1}{2}\, \sigma \models (\mathbf{1}_{X_>} \otimes \mathbf{1}) \;\&\; Eq^{\perp} \;+\; \tfrac{1}{2}\, \sigma \models (\mathbf{1} \otimes \mathbf{1}_{\neg X_>}) \;\&\; Eq^{\perp} \\ &\leq\; \tfrac{1}{2}\, \sigma \models Eq^{\perp} \;+\; \tfrac{1}{2}\, \sigma \models Eq^{\perp} \\ &=\; \sigma \models Eq^{\perp}.\end{aligned}$$

For the inequality $(\geq)$ one uses what is called an *optimal coupling* $\rho \in \mathcal{D}(X \times X)$ of ω_1, ω_2. It can be defined as:

$$\rho(x,y) \;:=\; \begin{cases} \min\big(\omega_1(x), \omega_2(x)\big) & \text{if } x = y \\ \dfrac{\max\big(\omega_1(x) - \omega_2(x), 0\big) \cdot \max\big(\omega_2(y) - \omega_1(y), 0\big)}{d(\omega_1,\omega_2)} & \text{otherwise.} \end{cases} \tag{4.17}$$

We first check that this ρ is a coupling. Let $x \in X_>$ so that $\omega_1(x) > \omega_2(x)$; then:

$$\begin{aligned}&\textstyle\sum_y \rho(x,y) \\ &=\; \omega_2(x) + (\omega_1(x) - \omega_2(x)) \cdot \sum_{y\neq x} \frac{\max\big(\omega_2(y) - \omega_1(y), 0\big)}{d(\omega_1,\omega_2)} \\ &=\; \omega_2(x) + (\omega_1(x) - \omega_2(x)) \cdot \frac{\sum_{y\in X_<} \omega_2(y) - \omega_1(y)}{d(\omega_1,\omega_2)} \\ &=\; \omega_2(x) + (\omega_1(x) - \omega_2(x)) \cdot \frac{X{\downarrow}}{d(\omega_1,\omega_2)} \\ &=\; \omega_2(x) + (\omega_1(x) - \omega_2(x)) \cdot 1 \qquad \text{see the proof of Proposition 4.5.1} \\ &=\; \omega_1(x).\end{aligned}$$

If $x \notin X_>$, so that $\omega_1(x) \leq \omega_2(x)$, then it is obvious that $\sum_y \rho(x,y) = \omega_1(x) + 0 = \omega_1(x)$. This shows $\sigma[1,0] = \omega_1$. In a similar way one obtains $\sigma[0,1] = \omega_2$. Finally,

$$\begin{aligned}\rho \models Eq \;=\; \textstyle\sum_x \rho(x,x) \;&=\; \textstyle\sum_x \min\big(\omega_1(x), \omega_2(x)\big) \\ &=\; \textstyle\sum_{x\in X_>} \omega_2(x) \;+\; \sum_{x\notin X_>} \omega_1(x) \\ &=\; \omega_2(X_>) + 1 - \omega_1(X_>) \\ &=\; 1 - \big(\omega_1(X_>) - \omega_2(X_>)\big) \;=\; 1 - d(\omega_1,\omega_2).\end{aligned}$$

Hence $d(\omega_1,\omega_2) = 1 - (\rho \models Eq) = \rho \models Eq^{\perp}$. ❑

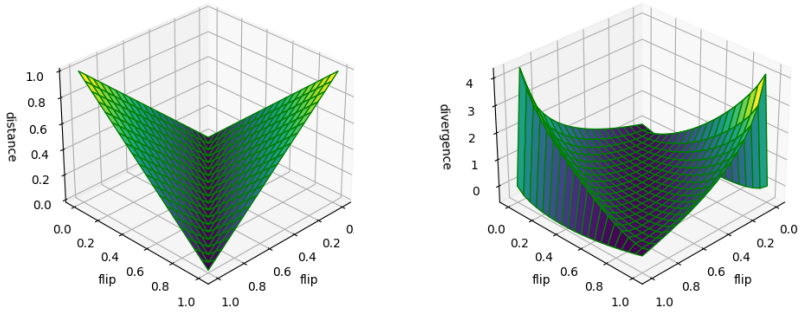


Figure 4.1 Visual comparance of distance and divergence between flip states, see Remark 4.5.5 for details.

Remark 4.5.5. In Definition 2.7.1 we have seen Kullback-Leibler divergence D_{KL}, as a measure of difference between states. However, this D_{KL} is not a proper metric, since it is not symmetric, see Exercise 2.7.1. Nevertheless, it is often used as distance between states, especially in minimisation problems (see *e.g.* Exercise **??**).

Figure 4.1 compares the total variation distance and the Kullback-Leibler divergence between two flip states:

$$d\Big(\mathit{flip}(r), \mathit{flip}(s)\Big) \qquad\qquad D_{KL}\Big(\mathit{flip}(r), \mathit{flip}(s)\Big).$$

for $r, s \in [0,1]$, where, recall, $\mathit{flip}(r) = r|1\rangle + (1-r)|0\rangle$. The distance and divergence are zero when $r = s$ and increases on both sides of the diagonal. The distance ascends via straight planes, but the divergence has a more baroque shape.

The relation between distributions and multisets is a recurring theme. Now that we have a distance function on distributions we can speak about approximation of a distribution via a chain of multisets. This is what the next remark is about. This topic returns as a law of large numbers in Section 5.5, see esp. Theorem 5.5.4. Here we take a an algorithmic perspective.

Remark 4.5.6. Let a distribution $\omega \in \mathcal{D}(X)$ be given. One can ask: is there a squence of natural multisets $\varphi_K \in \mathcal{N}[K](X)$ with $\mathit{Flrn}(\varphi_K)$ getting closer and closer, in the total variation distance d, as K goes to infinity?

The answer is yes. Here is one way to do it. Assume the distribution ω has

support $\{x_1, \ldots, x_n\}$, of course with $n > 0$. Define $\omega_1 := \mathbf{0}$, the empty multiset in $\mathcal{M}(\{x_1, \ldots, x_n\})$.

- Pick $\sigma_1 = 1|x_j\rangle$ where ω takes a maximum at x_j, *i.e.*, $\omega(x_j) \geq \omega(x_i)$ for all i. Set variable $cp := j$, for 'current position'.
- Set $\omega_{n+1} := \omega_n + \omega$, in $\mathcal{M}(X)$. Look for the first position i after cp where $\sigma_K(x_i) < \omega_{n+1}(x_i)$. Then you set $\sigma_{K+1} := \sigma_K + 1|x_i\rangle$, and $cp := i$. This search wraps around, if needed. When no i is found we are done and have $\mathit{Flrn}(\sigma_K) = \omega$.

Concretely, if $\omega = \frac{1}{6}|x_1\rangle + \frac{1}{2}|x_2\rangle + \frac{1}{3}|x_3\rangle$. Then, consecutively,

- $\omega_0 = \mathbf{0}$ and $\sigma_1 = 1|x_2\rangle$
- $\omega_1 = \frac{1}{6}|x_1\rangle + \frac{1}{2}|x_2\rangle + \frac{1}{3}|x_3\rangle$ and $\sigma_2 = 1|x_2\rangle + 1|x_3\rangle$
- $\omega_2 = \frac{1}{3}|x_1\rangle + 1|x_2\rangle + \frac{2}{3}|x_3\rangle$ and $\sigma_3 = 1|x_1\rangle + 1|x_2\rangle + 1|x_3\rangle$
- $\omega_3 = \frac{1}{2}|x_1\rangle + \frac{3}{2}|x_2\rangle + 1|x_3\rangle$ and $\sigma_4 = 1|x_1\rangle + 2|x_2\rangle + 1|x_3\rangle$
- $\omega_4 = \frac{2}{3}|x_1\rangle + 2|x_2\rangle + \frac{4}{3}|x_3\rangle$ and $\sigma_5 = 1|x_1\rangle + 2|x_2\rangle + 2|x_3\rangle$
- $\omega_5 = \frac{5}{6}|x_1\rangle + \frac{5}{2}|x_2\rangle + \frac{5}{3}|x_3\rangle$ and $\sigma_6 = 1|x_1\rangle + 3|x_2\rangle + 2|x_3\rangle$.

Indeed, $\mathit{Flrn}(\sigma_6) = \omega$. In this case we get a finite sequence of multisets approaching ω. The sequence is infinite for, *e.g.*,

$$\omega = \tfrac{1}{7}|x_1\rangle + \tfrac{4}{7}|x_2\rangle + \tfrac{2}{7}|x_3\rangle \approx 0.1429|x_1\rangle + 0.5714|x_2\rangle + 0.2857|x_3\rangle.$$

Running the above algorithm gives, for instance:

- $\varphi_{10} = 2|x_1\rangle + 5|x_2\rangle + 3|x_3\rangle$
- $\varphi_{100} = 15|x_1\rangle + 56|x_2\rangle + 29|x_3\rangle$
- $\varphi_{1000} = 143|x_1\rangle + 571|x_2\rangle + 286|x_3\rangle$
- $\varphi_{10000} = 1429|x_1\rangle + 5713|x_2\rangle + 2858|x_3\rangle$.

The metric on distributions is complete, when the sample space is finite. This is a standard result, see *e.g.* [86, Lemma 2.4]. Moreover, natural multisets can be used to get a dense subset.

Theorem 4.5.7. *1 If X is a finite set, then $\mathcal{D}(X)$, with the total variation distance d, is a complete metric space.*

2 For an arbitrary set X, the set $\mathcal{D}(X)$ has a dense subset:

$$\bigcup_{K \in \mathbb{N}} \mathcal{D}[K](X) \subseteq \mathcal{D}(X) \quad \textit{where} \quad \mathcal{D}[K](X) := \{\mathit{Flrn}(\varphi) \mid \varphi \in \mathcal{N}[K](X)\}.$$

This dense subset is countable when X is finite.

We often refer to the elements of $\mathcal{D}[K](X)$ as fractional distributions.

The combination of the these two points says that for a finite set X, the set $\mathcal{D}(X)$ of distributions on X is a Polish space: a complete metric space with a countable dense subset, see **??** for more details.

Proof. 1 Let $X = \{x_1, \ldots, x_\ell\}$ and $\omega_i \in \mathcal{D}(X)$ be a Cauchy sequence. Fix n. Then for all i, j,

$$\big|\, \omega_i(x_n) - \omega_j(x_n) \big| \;\leq\; 2 \cdot d(\omega_i, \omega_j).$$

Hence, the sequence $\omega_i(x_n) \in [0, 1]$ is a Cauchy sequence, say with limit $r_n \in [0, 1]$. Take $\omega = \sum_n r_n |x_n\rangle \in \mathcal{D}(X)$. This is the limit of the distributions ω_i.

2 Let $\omega \in \mathcal{D}(X)$ and $\varepsilon > 0$. We need to find a multiset $\varphi \in \mathcal{N}(X)$ with $d(\omega, \mathit{Flrn}(\varphi)) < \varepsilon$. There is a systematic way to find such multisets via the decimal representation of the probabilities in ω. This works as follows. Assume we have:

$$\omega \;=\; 0.383914217\ldots \big|a\big\rangle \;+\; 0.406475610\ldots \big|b\big\rangle \;+\; 0.209610173\ldots \big|c\big\rangle.$$

For each n we chop off after n decimals and multiply with 10^n, giving:

$$\begin{array}{lll}
\varphi_1 \coloneqq 3\big|a\big\rangle + 4\big|b\big\rangle + 2\big|c\big\rangle & \text{with} & d(\omega, \mathit{Flrn}(\varphi_1)) \leq \frac{1}{2} \cdot 3 \cdot 10^{-1} \\
\varphi_2 \coloneqq 38\big|a\big\rangle + 40\big|b\big\rangle + 20\big|c\big\rangle & \text{with} & d(\omega, \mathit{Flrn}(\varphi_2)) \leq \frac{1}{2} \cdot 3 \cdot 10^{-2} \\
\varphi_3 \coloneqq 383\big|a\big\rangle + 406\big|b\big\rangle + 209\big|c\big\rangle & \text{with} & d(\omega, \mathit{Flrn}(\varphi_3)) \leq \frac{1}{2} \cdot 3 \cdot 10^{-3} \\
\quad \textit{etc.}
\end{array}$$

In general, for a distribution ω with $\mathit{supp}(\omega) = \{x_1, \ldots, x_\ell\}$ we can thus construct a sequence of multisets $\varphi_n \in \mathcal{N}(\{x_1, \ldots, x_\ell\})$ with $d(\omega, \mathit{Flrn}(\varphi_n)) \leq \frac{1}{2} \cdot \ell \cdot 10^{-n}$. This distance becomes less than any given $\varepsilon > 0$, by choosing n sufficiently large.

When X is finite, say with n elements, we know from Proposition 1.7.4 that $\mathcal{N}[K](X)$ contains $\left(\!\binom{n}{K}\!\right)$ multisets, so that $\mathcal{D}[K](X)$ contains $\left(\!\binom{n}{K}\!\right)$ distributions. Hence a countable union $\bigcup_K \mathcal{D}[K](X)$ of such finite sets is countable. ❑

Notice that the sequence of multisets φ_K approaching ω described in Remark 4.5.6 has the special property that $\|\varphi_K\| = K$. This does not hold for the sequence φ_n in the above proof. Such a size property is not needed there.

The denseness of the fractional distributions is illustrated in Figure 4.2.

A different way to approximate a distribution ω is described in Section 5.5, via what is called the law of large numbers.

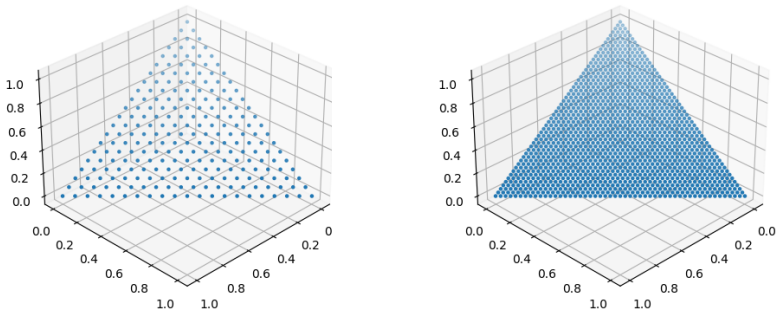


Figure 4.2 Fractional distributions from $\mathcal{D}[K](\{a, b, c\})$ as points in the cube $[0, 1]^3$, for $K = 20$ on the left and $K = 50$ on the right. The plot on the left contains $\left(\binom{3}{20}\right) = 231$ dots (distributions) and the one on the right $\left(\binom{3}{50}\right) = 1326$, see Proposition 1.7.4.

4.5.2 Distance between predicates

What we have to say about the validity-based distance (4.15) between predicates is rather brief. First, there is also a pointwise formulation.

Lemma 4.5.8. *For two predicates $p_1, p_2 \in \mathsf{Pred}(X)$,*

$$d(p_1, p_2) = \bigvee_{x \in X} \big| p_1(x) - p_2(x) \big|.$$

This distance function d makes the set $\mathsf{Pred}(X)$ into a metric space.

Proof. First, we have for each $x \in X$,

$$\begin{aligned} d(p_1, p_2) &\overset{(4.15)}{=} \bigvee_{\omega \in \mathcal{D}(X)} \big| \omega \models p_1 - \omega \models p_2 \big| \\ &\geq \big| \mathit{unit}(x) \models p_1 - \mathit{unit}(x) \models p_2 \big| \\ &= \big| p_1(x) - p_2(x) \big|. \end{aligned}$$

Hence $d(p_1, p_2) \geq \bigvee_x \big| p_1(x) - p_2(x) \big|$.

The other direction follows from:

$$\begin{aligned}\big|\omega \models p_1 \;-\; \omega \models p_2\big| &= \big|\textstyle\sum_z \omega(z)\cdot p_1(z) \;-\; \sum_z \omega(z)\cdot p_2(z)\big| \\ &\leq \textstyle\sum_z \omega(z)\cdot \big|p_1(z) \;-\; p_2(z)\big| \\ &\leq \textstyle\sum_z \omega(z)\cdot \bigvee_x \big|p_1(x) \;-\; p_2(x)\big|. \\ &= \big(\textstyle\sum_z \omega(z)\big)\cdot \bigvee_x \big|p_1(x) \;-\; p_2(x)\big|. \\ &= \textstyle\bigvee_x \big|p_1(x) \;-\; p_2(x)\big|.\end{aligned}$$

The fact that we get a metric space is now straightforward. ❑

There is an analogue of Lemma 4.5.3.

Lemma 4.5.9. *Predicate transformation is also non-expansive: for a channel $c\colon X \nrightarrow Y$ one has, for predicates $p_1, p_2 \in \mathcal{P}red(Y)$,*

$$d(c \mathrel{=\!\!\ll} p_1, c \mathrel{=\!\!\ll} p_2) \;\leq\; d(p_1, p_2).$$

Proof. Via the formulation of Lemma 4.5.8 we get:

$$\begin{aligned}d(c \mathrel{=\!\!\ll} p_1, c \mathrel{=\!\!\ll} p_2) &= \textstyle\bigvee_x \big|(c \mathrel{=\!\!\ll} p_1)(x) \;-\; (c \mathrel{=\!\!\ll} p_2)(x)\big| \\ &= \textstyle\bigvee_x \big|\sum_y c(x)(y)\cdot p_1(y) \;-\; \sum_y c(x)(y)\cdot p_2(y)\big| \\ &\leq \textstyle\bigvee_x \sum_y c(x)(y)\cdot \big|p_1(y) \;-\; p_2(y)\big| \\ &\leq \textstyle\bigvee_x \big(\sum_y c(x)(y)\big)\cdot d(p_1, p_2) \\ &= d(p_1, p_2).\end{aligned}$$

❑

Exercises

4.5.1 1 Prove that for states $\omega, \omega' \in \mathcal{D}(X)$ and $\rho, \rho' \in \mathcal{D}(Y)$ there is an inequality:

$$d(\omega \otimes \rho, \omega' \otimes \rho') \;\leq\; d(\omega, \omega') + d(\rho, \rho').$$

(In this situation thereis an equality for Kullback-Leibler divergence, see Lemma 2.7.2 (2).)

2 Prove similarly that for predicates $p, p' \in \mathcal{P}red(X)$ and $q, q' \in \mathcal{P}red(Y)$ one gets:

$$d(p \otimes q, p' \otimes q') \;\leq\; d(p, p') + d(q, q').$$

4.5.2 In the context of Remark 4.5.5, check that:

$$\begin{aligned}d\big(\mathit{flip}(0), \mathit{flip}(1)\big) \;&=\; 1 \;=\; d\big(\mathit{flip}(1), \mathit{flip}(0)\big) \\ D_{KL}\big(\mathit{flip}(0), \mathit{flip}(1)\big) \;&=\; 0 \;=\; D_{KL}\big(\mathit{flip}(1), \mathit{flip}(0)\big).\end{aligned}$$

4.5.3 In [88] the influence of a predicate p on a state ω is measured via the distance $d(\omega, \omega|_p)$. This influence can be zero, for the truth predicate $p = \mathbf{1}$.

Consider the set $\{H, T\}$ with state $\sigma(\varepsilon) = \varepsilon|H\rangle + (1-\varepsilon)|T\rangle$, and with predicate $p = \mathbf{1}_H$. Prove that $d(\sigma(\varepsilon), \sigma(\varepsilon)|_p) \to 1$ as $\varepsilon \to 0$.

4.5.4 This exercise uses the distance between a joint state and the product of its marginals as measure of entwinedness, like in [73].

1 Take $\sigma_2 \coloneqq \frac{1}{2}|00\rangle + \frac{1}{2}|11\rangle \in \mathcal{D}(2 \times 2)$, for $2 = \{0, 1\}$. Show that:

$$d\Big(\sigma_2,\ \sigma_2[1,0] \otimes \sigma_2[0,1]\Big) = \tfrac{1}{2}.$$

2 Take $\sigma_3 \coloneqq \frac{1}{2}|000\rangle + \frac{1}{2}|111\rangle \in \mathcal{D}(2 \times 2 \times 2)$. Show that:

$$d\Big(\sigma_3,\ \sigma_3[1,0,0] \otimes \sigma_3[0,1,0] \otimes \sigma_3[0,0,1]\Big) = \tfrac{3}{4}.$$

3 Now define $\sigma_n \in \mathcal{D}(2^n)$ for $n \geq 2$ as:

$$\sigma_n \coloneqq \tfrac{1}{2}|\underbrace{0\cdots0}_{n \text{ times}}\rangle + \tfrac{1}{2}|\underbrace{1\cdots1}_{n \text{ times}}\rangle.$$

Show that:

- each marginal $\pi_i \gg\!= \sigma_n$ equals $\frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle$;
- the product $\bigotimes_i (\pi_i \gg\!= \sigma_n)$ of the marginals is the uniform state on 2^n;
- $d\Big(\sigma_n,\ \bigotimes_i (\pi_i \gg\!= \sigma_n)\Big) = \dfrac{2^{n-1} - 1}{2^{n-1}}$.

Note that the latter distance goes to 1 as n goes to infinity.

4.5.5 Show that $\mathcal{D}(X)$, with the total variation distance (4.5), is a complete metric space when the set X is finite.

4.5.6 The next 'splitting lemma' is attributed to Jones [93], see *e.g.* [95, 117]. For $\omega_1, \omega_2 \in \mathcal{D}(X)$ with distance $d \coloneqq d(\omega_1, \omega_2)$ one can find distributions $\omega_1', \omega_2', \sigma \in \mathcal{D}(X)$ so that both ω_1 and ω_2 can be written as convex sum:

$$\omega_i = d \cdot \omega_i' + (1-d) \cdot \sigma.$$

Prove this result.

Hint: Use the optimal coupling ρ from (4.17) to define $\sigma(x) = \frac{\rho(x,x)}{1-d}$.

5

Variance and covariance

The previous chapter introduced validity $\omega \models p$, of an observable p in a state/distribution ω. The current chapter uses validity to define the standard statistical concepts of variance and covariance, and the associated notions of standard deviation and correlation. Informally, for a random variable (ω, p), the variance $\mathit{Var}(\omega, p)$ describes to which extent the observable p differs from the expected value $\omega \models p$, that is, how much much p varies or is spread out. Together, $\omega \models p$ and $\mathit{Var}(\omega, p)$ are representational values that capture the statistical essence of a random variable. The standard deviation of a random variable is the square root of its variance.

The notion of covariance is used to compare *two* random variables. What do we mean by *two*? We can have:

1 two random variables (ω, p_1) and (ω, p_2), with (possibly) different observables $p_1, p_2 \colon X \to \mathbb{R}$, but with the same shared state $\omega \in \mathcal{D}(X)$;
2 a joint state $\tau \in \mathcal{D}(X_1 \times X_2)$ together with two observables $q_1 \colon X_1 \to \mathbb{R}$ and $q_2 \colon X_2 \to \mathbb{R}$ on the two components X_1, X_2. Via weakening of the observables we get two random variables:

$$(\tau, q_1 \otimes \mathbf{1}) \qquad\qquad (\tau, \mathbf{1} \otimes q_2)$$

like in the previous point, involving two observables $q_1 \otimes \mathbf{1}$ and $\mathbf{1} \otimes q_2$, now on the same set $X_1 \times X_2$.

These differences are significant, but the two cases are not always clearly distinguished in the literature. One of the principles in this book is to make states explicit. Hence we shall clearly distinguish between the first *shared-state* form of covariance and the second *joint-state* form.

Apart from that, covariance captures to what extent two random variables change together. Covariance may be positive, when the variables change together, or negative, meaning that they change in opposite directions.

This short chapter first introduces the basic definitions and results for variance, and for covariance in shared-state form. They are applied to draw distributions, from Chapter 3, in Section 5.2. The joint-state version of covariance is introduced in Section 5.3 and illustrated in several examples. Section 5.4 then establishes the equivalence between:

- non-entwinedness of a joint state, meaning that it is the product of its marginals;
- joint-state independence of random variables on this state
- joint-state covariance is zero.

See Theorem 5.4.6 for details. Such equivalences do not hold for shared-state formulations. This is the very reason for being careful about the distinction between a shared state and a joint state.

Covariance and correlation of (observables on) joint states is relevant in the setting of updating, in the next chapter. In presence of such correlation, updating in one (product) component has crossover influence in the other component.

At the end of this chapter we use what we have learned about variance, to formulate what is called the weak law of large numbers. It shows that by accumulating repeated draws from a distribution one comes arbitrary close to that distribution. This is an alternative way of expressing the denseness of fractional distributions in among all distributions, as formulated in Theorem 4.5.7.

5.1 Variance and shared-state covariance

This section describes the standard notions of variance, covariance and correlation within the setting of this book. It uses the validity relation $\models$ and the operations on observables from Section 4.2. Recall that we understand a random variable here as a pair (ω, p) consisting of a state $\omega \in \mathcal{D}(X)$ and an observable $p\colon X \to \mathbb{R}$. The validity $\omega \models p$ is a real number, and can thus be used as a scalar, in the sense of Section 4.2. The truth predicate forms an observable $\mathbf{1} \in \mathit{Obs}(X)$; scalar multiplication yields a new observable $(\omega \models p)\cdot\mathbf{1} \in \mathit{Obs}(X)$. It can be subtracted[1] from p, and then squared, giving an observable:

$$(p - (\omega \models p) \cdot \mathbf{1})^2 \;=\; (p - (\omega \models p) \cdot \mathbf{1}) \mathbin{\&} (p - (\omega \models p) \cdot \mathbf{1}) \;\in\; \mathit{Obs}(X).$$

This observable denotes the function that sends $x \in X$ to $(p(x) - (\omega \models p))^2 \in$

[1] Subtraction expressions like these occur more frequently in mathematics. For instance, an eigenvalue λ of a matrix M may be defined as the scalar that forms a solution to the equation $M - \lambda \cdot \mathbf{1} = 0$, where $\mathbf{1}$ is the identity matrix. A similar expression is used to define the elements in the spectrum of a C^*-algebra. See also Excercise 4.2.12.

$\mathbb{R}_{\geq 0}$. It is thus a factor. Its validity in the original state ω is called variance. It captures how far the values of p are spread out from their expected value.

Definition 5.1.1. For a random variable (ω, p), the *variance* $\mathit{Var}(\omega, p)$ is the non-negative number defined by:

$$\mathit{Var}(\omega, p) \;:=\; \omega \models (p - (\omega \models p) \cdot \mathbf{1})^2.$$

When the underlying sample space X is a subset of $\mathbb{R}$, say via $\mathit{incl} \colon X \hookrightarrow \mathbb{R}$, we simply write $\mathit{Var}(\omega)$ for $\mathit{Var}(\omega, \mathit{incl})$.

The name *standard deviation* is used for the square root of the variance; thus:

$$\mathit{StDev}(\omega, p) \;:=\; \sqrt{\mathit{Var}(\omega, p)}.$$

Example 5.1.2. 1 We recall Example 4.1.4 (1), with distribution $\mathit{flip}(\frac{3}{10}) = \frac{3}{10}|1\rangle + \frac{7}{10}|0\rangle$ and observable $v(0) = -50$ and $v(1) = 100$. We had $\omega \models v = -5$, and so we get:

$$\begin{aligned}\mathit{Var}\Big(\mathit{flip}(\tfrac{3}{10}), v\Big) &= \sum_{x \in \{0,1\}} \mathit{flip}(\frac{3}{10})(x) \cdot (v(x) + 5)^2 \\ &= \tfrac{3}{10} \cdot (100 + 5)^2 + \tfrac{7}{10} \cdot (-50 + 5)^2 \;=\; 4725.\end{aligned}$$

The standard deviation is around 68.7.

2 For a (fair) dice we have $\mathit{pips} = \{1, 2, 3, 4, 5, 6\} \hookrightarrow \mathbb{R}$ and $\mathit{mean}(\mathit{dice}) = \frac{7}{2}$ so that:

$$\begin{aligned}\mathit{Var}(\mathit{dice}) &= \sum_{x \in \mathit{pips}} \mathit{dice}(x) \cdot (x - \tfrac{7}{2})^2 \\ &= \tfrac{1}{6} \cdot \Big((\tfrac{5}{2})^2 + (\tfrac{3}{2})^2 + (\tfrac{1}{2})^2 + (\tfrac{1}{2})^2 + (\tfrac{3}{2})^2 + (\tfrac{5}{2})^2\Big) \;=\; \tfrac{35}{12}.\end{aligned}$$

Via a suitable shift-and-rescale one can standardise an observable so that its validity becomes 0 and its variance becomes 1, see Exercise 5.1.6.

The following result is often useful in calculations.

Lemma 5.1.3. *Variance satisfies:*

$$\mathit{Var}(\omega, p) \;=\; (\omega \models p^2) - (\omega \models p)^2.$$

Since variance is non-negative, we get as byproduct an inequality:

$$\omega \models p^2 \;\geq\; (\omega \models p)^2.$$

Proof. We have:

$$\begin{aligned}
&\mathit{Var}(\omega, p)\\
&= \omega \models (p - (\omega \models p) \cdot \mathbf{1})^2\\
&= \textstyle\sum_x \omega(x)(p(x) - (\omega \models p))^2\\
&= \textstyle\sum_x \omega(x)\Big(p(x)^2 - 2(\omega \models p)p(x) + (\omega \models p)^2\Big)\\
&= \Big(\textstyle\sum_x \omega(x)p^2(x)\Big) - 2(\omega \models p)\Big(\textstyle\sum_x \omega(x)p(x)\Big) + \Big(\textstyle\sum_x \omega(x)(\omega \models p)^2\Big)\\
&= (\omega \models p^2) - 2(\omega \models p)(\omega \models p) + (\omega \models p)^2\\
&= (\omega \models p^2) - (\omega \models p)^2.
\end{aligned}$$

❑

This result is used to obtain the variances of a draw distributions in a later section.

We continue with covariance and correlation, which involve *two* random variables, instead one, as for variance. We can distinguish two forms of these, namely:

- The two random variables are of the form (ω, p_1) and (ω, p_2), where they share their state ω.
- There is a joint state $\tau \in \mathcal{D}(X_1 \times X_2)$ together with two observables $q_1 : X_1 \to \mathbb{R}$ and $q_2 : X_2 \to \mathbb{R}$ on the two components X_1, X_2. This situation can be seen as a special case of the previous point by first weakening the two observables to the product space, via: $\pi_1 =\!\ll q_1 = q_1 \otimes \mathbf{1}$ and $\pi_2 =\!\ll q_2 = \mathbf{1} \otimes q_2$. Thus we obtain two random variable with a shared state:

  $$(\tau,\ \pi_1 =\!\ll q_1) \qquad \text{and} \qquad (\tau,\ \pi_2 =\!\ll q_2).$$

  These observable transformations $\pi_i =\!\ll q_i$ along a deterministic channel $\pi_i : X_1 \times X_2 \to X_i$ can also be described simply as function composition $q_i \circ \pi_i : X_1 \times X_2 \to \mathbb{R}$, see Lemma 4.3.2 (7).

We start with the situation in the first bullet above, and deal with the second bullet in Definition 5.3.1 in Section 5.3.

Definition 5.1.4. Let (ω, p_1) and (ω, p_2) be two random variable with a shared state $\omega \in \mathcal{D}(X)$.

1 The *covariance* of these random variables is defined as the validity:

$$\mathit{Cov}(\omega, p_1, p_2) \ :=\ \omega \models (p_1 - (\omega \models p_1) \cdot \mathbf{1}) \mathbin{\&} (p_2 - (\omega \models p_2) \cdot \mathbf{1}).$$

2 The *correlation* between (ω, p_1) and (ω, p_2) is the covariance divided by their standard deviations:

$$\mathit{Cor}(\omega, p_1, p_2) \ :=\ \frac{\mathit{Cov}(\omega, p_1, p_2)}{\mathit{StDev}(\omega, p_1) \cdot \mathit{StDev}(\omega, p_2)}.$$

Notice that variance $\mathit{Var}(\omega, p)$ is a special case of covariance $\mathit{Cov}(\omega, p, p)$, namely with equal observables. Hence if there is an inclusion $\mathit{incl}\colon X \hookrightarrow \mathbb{R}$ and we would use this inclusion twice to compute covariance, we are in fact computing variance.

Correlation is normalised covariance, so that the outcome is in the interval $[-1, 1]$, see Exercise 5.1.7 below. In ordinary language two phenomena are called correlated when there is relation between them. More technically, two random variables are called correlated if their correlation, as defined above, is non-zero — or equivalently, when their covariance is non-zero. Positive correlation means that the observables mover together in the same direction, whereas negative correlation means that they move in opposite directions.

Before we go on, we state the following analogue of Lemma 5.1.3, leaving the proof to the reader.

Lemma 5.1.5. *Covariance can be reformulated as:*

$$\mathit{Cov}(\omega, p_1, p_2) = (\omega \models p_1 \mathbin{\&} p_2) - (\omega \models p_1) \cdot (\omega \models p_2). \qquad \blacksquare$$

Example 5.1.6. 1 We have seen in Definition 4.1.2 (2) how the average of an observable can be computed as its validity in a uniform state. The same approach is used to compute the covariance (and correlation) in a uniform joint state. Consider the following to lists a and b of numerical data, of the same length.

$$a = [5, 10, 15, 20, 25] \qquad\qquad b = [10, 8, 10, 15, 12]$$

We will identify a and b with random variables, namely with $a, b\colon 5 \to \mathbb{R}$, where $5 = \{0, 1, 2, 3, 4\}$. Hence there are obvious definitions: $a(0) = 5$, $a(1) = 10$, $a(2) = 15$, $a(3) = 20$, $a(4) = 25$, and similiarly for b. Then we can compute their averages as validities in the uniform state unif_5 on the set 5:

$$\mathit{avg}(a) = \mathit{unif}_5 \models a = 15 \qquad\qquad \mathit{avg}(a) = \mathit{unif}_5 \models a = 11.$$

We will calculate the covariance between a and b wrt. the uniform state unif_5, as:

$$\begin{aligned} \mathit{Cov}(\mathit{unif}_5, a, b) &= \mathit{unif}_5 \models (a - (\mathit{unif}_5 \models a) \cdot \mathbf{1}) \mathbin{\&} (b - (\mathit{unif}_5 \models b) \cdot \mathbf{1}) \\ &= \textstyle\sum_i \frac{1}{5} \cdot (a(i) - 15) \cdot (b(i) - 11) \\ &= 11. \end{aligned}$$

2 In order to obtain the correlation between a, b, we first need to compute their

variances:

$$\begin{aligned} Var(unif_5, a) &= \textstyle\sum_i \frac{1}{5}(a(i) - 15)^2 = 50 \\ Var(unif_5, b) &= \textstyle\sum_i \frac{1}{5}(b(i) - 11)^2 = 5.6 \end{aligned}$$

Then:

$$Cor(unif_5, a, b) = \frac{Cov(unif_5, a, b)}{\sqrt{Var(unif_5, a)} \cdot \sqrt{Var(unif_5, b)}} = \frac{11}{\sqrt{50} \cdot \sqrt{5.6}} \approx 0.66.$$

The next result collects several linearity properties for (co)variance and correlation. This shows that one can do a lot of re-scaling and stretching of observables without changing the outcome.

Theorem 5.1.7. *Consider a state $\omega \in \mathcal{D}(X)$, with observables $p, p_1, p_2 \in Obs(X)$ and numbers $r, s \in \mathbb{R}$.*

1 Covariance satisfies:

$$\begin{aligned} Cov(\omega, p_1, p_2) &= Cov(\omega, p_2, p_1) \\ Cov(\omega, p_1, \mathbf{1}) &= 0 \\ Cov(\omega, r \cdot p_1, p_2) &= r \cdot Cov(\omega, p_1, p_2) \\ Cov(\omega, p, p_1 + p_2) &= Cov(\omega, p, p_1) + Cov(\omega, p, p_2) \\ Cov(\omega, p_1 + r \cdot \mathbf{1}, p_2 + s \cdot \mathbf{1}) &= Cov(\omega, p_1, p_2). \end{aligned}$$

2 Variance satisfies:

$$\begin{aligned} Var(\omega, r \cdot p) &= r^2 \cdot Var(\omega, p) \\ Var(\omega, p + r \cdot \mathbf{1}) &= Var(\omega, p) \\ Var(\omega, p_1 + p_2) &= Var(\omega, p_1) + 2 \cdot Cov(\omega, p_1, p_2) + Var(\omega, p_2). \end{aligned}$$

3 For correlation we have:

$$\begin{aligned} Cor(\omega, p_1, p_2) &= Cor(\omega, p_2, p_1) \\ Cor(\omega, r \cdot p_1, s \cdot p_2) &= \begin{cases} Cor(\omega, p_1, p_2) & \text{if } r, s \text{ have the same sign} \\ -Cor(\omega, p_1, p_2) & \text{otherwise.} \end{cases} \\ Cor(\omega, p_1 + r \cdot \mathbf{1}, p_2 + s \cdot \mathbf{1}) &= Cor(\omega, p_1, p_2). \end{aligned}$$

Proof. 1 Obviously, covariance is symmetric and covariance with truth is 0. Covariance preserves scalar multiplication in each (observable) argument since by Lemma 5.1.5:

$$\begin{aligned} Cov(\omega, r \cdot p_1, p_2) &= (\omega \models (r \cdot p_1) \mathbin{\&} p_2) - (\omega \models r \cdot p_1) \cdot (\omega \models p_2) \\ &= r \cdot (\omega \models p_1 \mathbin{\&} p_2) - r \cdot (\omega \models p_1) \cdot (\omega \models p_2) \\ &= r \cdot Cov(\omega, p_1, p_2). \end{aligned}$$

For preservation of sums we reason from the definition:

$$\begin{aligned}
&Cov(\omega, p, p_1 + p_2)\\
&= (\omega \models p \;\&\; (p_1 + p_2)) - (\omega \models p) \cdot (\omega \models p_1 + p_2)\\
&= (\omega \models (p \;\&\; p_1) + (p \;\&\; p_2)) - (\omega \models p) \cdot ((\omega \models p_1) + (\omega \models p_2))\\
&= (\omega \models p \;\&\; p_1) + (\omega \models p \;\&\; p_2)\\
&\qquad - (\omega \models p) \cdot (\omega \models p_1) - (\omega \models p) \cdot (\omega \models p_2)\\
&= Cov(\omega, p, p_1) + Cov(\omega, p, p_2)
\end{aligned}$$

The equation $Cov(\omega, p_1 + r \cdot \mathbf{1}, p_2 + s \cdot \mathbf{1}) = Cov(\omega, p_1, p_2)$ follows from the previous equations.

2 The first property holds by what we have just seen:

$$Var(\omega, r \cdot p) = Cov(\omega, r \cdot p, r \cdot p) = r^2 \cdot Cov(\omega, p, p) = r^2 \cdot Var(\omega, p).$$

Similarly, $Var(\omega, p + r \cdot \mathbf{1}) = Var(\omega, p)$. Next:

$$\begin{aligned}
&Var(\omega, p_1 + p_2)\\
&= Cov(\omega, p_1 + p_2, p_1 + p_2)\\
&= Cov(\omega, p_1 + p_2, p_1) + Cov(\omega, p_1 + p_2, p_2)\\
&= Cov(\omega, p_1, p_1) + Cov(\omega, p_2, p_1) + Cov(\omega, p_1, p_2) + Cov(\omega, p_2, p_2)\\
&= Var(\omega, p_1) + 2 \cdot Cov(\omega, p_1, p_2) + Var(\omega, p_2).
\end{aligned}$$

3 Symmetry of correlation is obvious. By unpacking the definition of correlation and using the previous two items we get:

$$\begin{aligned}
Cor(\omega, r \cdot p_1, s \cdot p_2) &= \frac{Cov(\omega, r \cdot p_1, s \cdot p_2)}{\sqrt{Var(\omega, r \cdot p_1)} \cdot \sqrt{Var(\omega, s \cdot p_2)}}\\
&= \frac{r \cdot s \cdot Cov(\omega, p_1, p_2)}{\sqrt{r^2 \cdot Var(\omega, p_1)} \cdot \sqrt{s^2 \cdot Var(\omega, p_2)}}\\
&= \frac{r \cdot s \cdot Cov(\omega, p_1, p_2)}{|r| \cdot \sqrt{Var(\omega, p_1)} \cdot |s| \cdot \sqrt{Var(\omega, p_2)}}\\
&= \begin{cases} Cor(\omega, p_1, p_2) & \text{if } r, s \text{ have the same sign} \\ -Cor(\omega, p_1, p_2) & \text{otherwise.} \end{cases}
\end{aligned}$$

(The same sign means: either both $r \geq 0$ and $s \geq 0$ or both $r \leq 0$ and $s \leq 0$.)

The final equation $Cor(\omega, p_1 + r \cdot \mathbf{1}, p_2 + s \cdot \mathbf{1}) = Cor(\omega, p_1, p_2)$ holds since both variance and covariance are closed under addition of constants. ❑

Exercises

5.1.1 Let ω be a state and p be a factor on the same set. Define for $v \in \mathbb{R}_{\geq 0}$,

$$f(v) \coloneqq \omega \models (p - v \cdot \mathbf{1})^2.$$

Show that the function $f \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ takes its minimum value at $\omega \models p$.

5.1.2 Let $\tau \in \mathcal{D}(X \times Y)$ be a joint state with an observable $p \colon X \to \mathbb{R}$. We can turn them into a random variable in two ways, by marginalisation and weakening:

$$(\tau[1, 0], p) \qquad \text{and} \qquad (\tau, \pi_1 \mathrel{=\!\!\ll} p),$$

where $\pi_1 \mathrel{=\!\!\ll} p = p \otimes \mathbf{1}$ is an observable on $X \times Y$.

These two random variables have the same expected value, by (4.8). Show that they have the same variance too:

$$\mathit{Var}(\tau[1, 0], p) = \mathit{Var}(\tau, \pi_1 \mathrel{=\!\!\ll} p).$$

As a result, the standard deviations are also the same.

5.1.3 Prove Lemma 5.1.5 along the lines of the proof of Lemma 5.1.3.

5.1.4 Show for a predicate p,

1 $\mathit{Cov}(\omega, p, p^\perp) = -\mathit{Var}(\omega, p) \leq 0$;
2 $\mathit{Var}(\omega, p^\perp) = \mathit{Var}(\omega, p)$.

5.1.5 Let $h \colon X \to Y$ be a function, with a state $\omega \in \mathcal{D}(X)$ on its domain an an observable $q \colon Y \to \mathbb{R}$ on its codomain. Show that:

$$\mathit{Var}\Big(\omega,\, q \circ h\Big) = \mathit{Var}\Big(\mathcal{D}(h)(\omega),\, q\Big).$$

Hint: Recall Exercise 4.3.3.

5.1.6 Let (ω, p) be a random variable on a space X. Define a new *standard score* observable $\mathit{StSc}(\omega, p) \colon X \to \mathbb{R}$ by:

$$\mathit{StSc}(\omega, p)(x) \coloneqq \frac{p(x) - (\omega \models p)}{\mathit{StDev}(\omega, p)}.$$

This pair of ω with $\mathit{StSc}(\omega, p)(x)$ is also called the Z-random variable. It is normalised in the sense that:

1 $\omega \models \mathit{StSc}(\omega, p) = 0$;
2 $\mathit{Var}(\omega, \mathit{StSc}(\omega, p)) = \mathit{StDev}(\omega, \mathit{StSc}(\omega, p)) = 1$.

Prove these two items.

5.1.7 Recall the Cauchy-Schwarz inequality, for real numbers $a_i, b_i \in \mathbb{R}$,

$$\Big(\sum_i a_i b_i\Big)^2 \leq \Big(\sum_i a_i^2\Big) \cdot \Big(\sum_i b_i^2\Big).$$

Use this inequality to prove that correlation is in the interval $[-1, 1]$.

5.1.8 Let $\omega \in \mathcal{D}(X)$ be a state with an n-test $\vec{p} = p_1, \ldots, p_n$, see Definition 4.2.2.

1 Define the (symmetric) covariance matrix as:

$$\mathit{CovMat}(\omega, \vec{p}) := \begin{pmatrix} \mathit{Cov}(\omega, p_1, p_1) & \cdots & \mathit{Cov}(\omega, p_1, p_n) \\ \vdots & & \vdots \\ \mathit{Cov}(\omega, p_n, p_1) & \cdots & \mathit{Cov}(\omega, p_n, p_n) \end{pmatrix}$$

Prove that all rows and all colums add up to 0 and that the entries on the diagonal are non-negative and have a sum below 1.

2 Next consider the vector v of validities and the symmetric matrix A of conjunctions:

$$v := \begin{pmatrix} \omega \models p_1 \\ \vdots \\ \omega \models p_n \end{pmatrix} \qquad A := \begin{pmatrix} \omega \models p_1 \,\&\, p_1 & \cdots & \omega \models p_1 \,\&\, p_n \\ \vdots & & \vdots \\ \omega \models p_n \,\&\, p_1 & \cdots & \omega \models p_n \,\&\, p_n \end{pmatrix}$$

Check that $\mathit{CovMat}(\omega, \vec{p}) = A - v \cdot v^T$, where $(-)^T$ is transpose.

5.1.9 In linear regression a finite collection $(a_i, b_i)_{1 \leq i \leq n}$ of real numbers $a_i, b_i \in \mathbb{R}$ is given. The aim is to find coefficients $v, w \in \mathbb{R}$ of a line $y = vx + w$ that best approximates these points. The error that is minimised is the 'sum of squared residuals' given as:

$$f(v, w) := \sum\nolimits_i \Big(b_i - (va_i + w)\Big)^2.$$

We redescribe the a_i, b_i as observables $a, b \colon \{1, 2, \ldots, n\} \to \mathbb{R}$, with $a(i) = a_i, b(i) = b_i$, together with the uniform distribution *unif* on the space $\{1, 2, \ldots, n\}$. Thus we have two random variables (unif, a) and (unif, b) with a shared state. Write:

$$\overline{a} := \tfrac{1}{n} \textstyle\sum_i a_i \qquad \overline{b} := \tfrac{1}{n} \textstyle\sum_i b_i.$$

By taking partial derivatives $\frac{\partial f}{\partial v}$ and $\frac{\partial f}{\partial w}$, setting them to zero, and using some elementary calculus, one obtains the best linear approximation of the (a_i, b_i) via coefficients given by the familiar formulas:

$$\hat{v} = \frac{\sum_i a_i(b_i - \overline{b})}{\sum_i a_i(a_i - \overline{a})} \qquad \hat{w} = \overline{b} - \hat{v}\,\overline{a}.$$

1 Derive the above formulas for $\hat{v}$ and $\hat{w}$.

2 Show that one can also write the slope $\hat{v}$ of the best line as:

$$\hat{v} \;=\; \frac{\sum_i (a_i - \overline{a})(b_i - \overline{b})}{\sum_i (a_i - \overline{a})^2}.$$

3 Check that this yields:

$$\hat{v} \;=\; \frac{\mathit{Cov}(\mathit{unif}, a, b)}{\mathit{Var}(\mathit{unif}, a)}.$$

4 Thus, with the above values $\hat{v}, \hat{w}$ the sum of squares of $b-(\hat{v}\cdot a+\hat{w}\cdot \mathbf{1})$ is minimal. Show that the latter expression can also be written in terms of standard scores, see Exercise 5.1.6, namely as:

$$b - (\hat{v} \cdot a + \hat{w} \cdot \mathbf{1}) \;=\; \mathit{StDev}(b) \cdot \Big(\mathit{StSc}(b) - \mathit{Cor}(a, b) \cdot \mathit{StSc}(a)\Big),$$

where we have omitted the uniform distribution for convenience. The right-hand-side shows that by using correlation as scalar one can bring the standard score of a closest to the standard score of b.

Linear regression is described here as a technique for obtaining the 'best' line, from data points (a_i, b_i). Once this line is found, one can use it for *prediction*: if we have an arbitrary first coordinate a we can predict the corresponding second coordinate as $\hat{v} \cdot a + \hat{w}$. For instance, if a_i is number of hours spent learning by student i, and b_i is the resulting mark of student i, then linear regression may give a reasonable prediction of the mark given a (new) number a of time spent on learning. Chapter **??** is devoted to learning techniques, of which linear regression is simple instance.

5.2 Draw distributions and their (co)variances

This section establishes standard (co)variance results for the multinomial, hypergeometric and Pólya draw distributions.

With results that we have collected so far it is easy to compute point-variances of the draw distributions, using the point-evaluation observables π_y from (4.10).

Proposition 5.2.1. *Let X be a set with a distinguished element $y \in X$ and with a number $K \in \mathbb{N}$.*

1 For an urn-distribution $\omega \in \mathcal{D}(X)$,

$$\mathit{Var}\Big(\mathit{mn}[K](\omega),\, \pi_y\Big) \;=\; K \cdot \omega(y) \cdot (1 - \omega(y)).$$

2 For a non-empty urn $\psi \in \mathcal{N}[L](X)$ of size $L \geq K$,

$$Var\Big(hg[K](\psi),\, \pi_y\Big) \;=\; K \cdot \frac{L-K}{L-1} \cdot Flrn(\psi)(y) \cdot (1 - Flrn(\psi)(y)).$$

3 Similarly, in the Pólya case we have:

$$Var\Big(pl[K](\psi),\, \pi_y\Big) \;=\; K \cdot \frac{L+K}{L+1} \cdot Flrn(\psi)(y) \cdot (1 - Flrn(\psi)(y)).$$

Proof. 1 Using the variance formula of Lemma 5.1.3 we get:

$$\begin{aligned}
&Var(mn[K](\omega),\, \pi_y) \\
&= (mn[K](\omega) \models \pi_y \;\&\; \pi_y) - (mn[K](\omega) \models \pi_y)^2 \\
&= \left(\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y)^2 \right) - \left(\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \right)^2 \\
&= K \cdot (K-1) \cdot \omega(y)^2 + K \cdot \omega(y) - \big(K \cdot \omega(y)\big)^2 \\
&\qquad \text{by Exercise 3.3.7 (3) and Lemma 3.3.4} \\
&= K \cdot (\omega(y) - \omega(y)^2) \;=\; K \cdot \omega(y) \cdot (1 - \omega(y)).
\end{aligned}$$

2 We now use Exercise 3.4.7 and Lemma 3.4.4 in:

$$\begin{aligned}
&Var(hg[K](\psi),\, \pi_y) \\
&= \left(\sum_{\varphi \in \mathcal{N}[K](X)} hg[K](\psi)(\varphi) \cdot \varphi(y)^2 \right) - \left(\sum_{\varphi \in \mathcal{N}[K](X)} hg[K](\psi)(\varphi) \cdot \varphi(y) \right)^2 \\
&= K \cdot Flrn(\psi)(y) \cdot \frac{(K-1) \cdot \psi(y) + (L-K)}{L-1} - \big(K \cdot Flrn(\psi)(y)\big)^2 \\
&= K \cdot Flrn(\psi)(y) \cdot \left(\frac{L(K-1) \cdot \psi(y) + L(L-K)}{L(L-1)} - \frac{(L-1)K \cdot \psi(y)}{L(L-1)} \right) \\
&= K \cdot Flrn(\psi)(y) \cdot \frac{(L-K) \cdot (L - \psi(y))}{L(L-1)} \\
&= K \cdot \frac{L-K}{L-1} \cdot Flrn(\psi)(y) \cdot (1 - Flrn(\psi)(y)).
\end{aligned}$$

3 Similarly:

$$\begin{aligned}
&Var(pl[K](\psi),\, \pi_y) \\
&= \left(\sum_{\varphi \in \mathcal{N}[K](X)} pl[K](\psi)(\varphi) \cdot \varphi(y)^2 \right) - \left(\sum_{\varphi \in \mathcal{N}[K](X)} pl[K](\psi)(\varphi) \cdot \varphi(y) \right)^2 \\
&= K \cdot Flrn(\psi)(y) \cdot \frac{(K-1) \cdot \psi(y) + (L+K)}{L+1} - \big(K \cdot Flrn(\psi)(y)\big)^2 \\
&= K \cdot Flrn(\psi)(y) \cdot \left(\frac{L(K-1) \cdot \psi(y) + L(L+K)}{L(L+1)} - \frac{(L+1)K \cdot \psi(y)}{L(L+1)} \right) \\
&= K \cdot Flrn(\psi)(y) \cdot \frac{(L+K) \cdot (L - \psi(y))}{L(L+1)} \\
&= K \cdot \frac{L+K}{L+1} \cdot Flrn(\psi)(y) \cdot (1 - Flrn(\psi)(y)).
\end{aligned}$$

❑

With these pointwise results we can turn variances into multisets, like in Definition 4.4.1. In the literature one finds descriptions of such variances as vectors but they presuppose an ordering on the points of the underlying space. The multiset description given below does not require such an ordering.

Definition 5.2.2. Let X be a set (of colours). The variances of the three draw distributions are defined as multisets over X in:

$$\begin{aligned}
Var\big(mn[K](\omega)\big) &:= \sum_{y \in X} Var(mn[K](\omega), \pi_y) \,\big|y\big\rangle \\
&= K \cdot \sum_{y \in X} \omega(y) \cdot (1 - \omega(y)) \,\big|y\big\rangle \\
Var\big(hg[K](\psi)\big) &:= \sum_{y \in X} Var(hg[K](\psi), \pi_y) \,\big|y\big\rangle \\
&= K \cdot \frac{L-K}{L-1} \cdot \sum_{y \in X} Flrn(\psi)(y) \cdot (1 - Flrn(\psi)(y)) \,\big|y\big\rangle \\
Var\big(pl[K](\psi)\big) &:= \sum_{y \in X} Var(pl[K](\psi), \pi_y) \,\big|y\big\rangle \\
&= K \cdot \frac{L+K}{L+1} \cdot \sum_{y \in X} Flrn(\psi)(y) \cdot (1 - Flrn(\psi)(y)) \,\big|y\big\rangle.
\end{aligned}$$

There are anologous results and definitions for covariance.

Proposition 5.2.3. *Let X be a set with a two different elements $y \neq z$ from X and with a number $K \in \mathbb{N}$.*

1 For an urn-distribution $\omega \in \mathcal{D}(X)$,

$$Cov\big(mn[K](\omega),\, \pi_y,\, \pi_z\big) = -K \cdot \omega(y) \cdot \omega(z).$$

2 For an urn $\psi \in \mathcal{N}[L](X)$ of size $L \geq K$,

$$Cov\Big(hg[K](\psi),\, \pi_y,\, \pi_z\Big) \;=\; -K \cdot \frac{L-K}{L-1} \cdot Flrn(\psi)(y) \cdot Flrn(\psi)(z).$$

3 In the Pólya case:

$$Cov\Big(pl[K](\psi),\, \pi_y,\, \pi_z\Big) \;=\; -K \cdot \frac{L+K}{L+1} \cdot Flrn(\psi)(y) \cdot Flrn(\psi)(z).$$

Proof. 1 Using the covariance formula of Lemma 5.1.5 and Exercise 3.3.7 we get:

$$\begin{aligned}
&Cov(mn[K](\omega),\, \pi_y,\, \pi_z)\\
&= (mn[K](\omega) \models \pi_y \;\&\; \pi_z) - (mn[K](\omega) \models \pi_y) \cdot (mn[K](\omega) \models \pi_z)\\
&= K \cdot (K-1) \cdot \omega(y) \cdot \omega(z) - K \cdot \omega(y) \cdot K \cdot \omega(z)\\
&= -K \cdot \omega(y) \cdot \omega(z).
\end{aligned}$$

2 In the same way, using Exercise 3.4.7 and Lemma 3.4.4:

$$\begin{aligned}
&Cov(hg[K](\psi),\, \pi_y,\, \pi_z)\\
&= (hg[K](\psi) \models \pi_y \;\&\; \pi_z) - (hg[K](\psi) \models \pi_y) \cdot (hg[K](\psi) \models \pi_z)\\
&= K \cdot (K-1) \cdot Flrn(\psi)(y) \cdot \frac{\psi(z)}{L-1} - K \cdot Flrn(\psi)(y) \cdot K \cdot Flrn(\psi)(z)\\
&= K \cdot Flrn(\psi)(y) \cdot \frac{L(K-1) \cdot \psi(z) - K(L-1) \cdot \psi(z)}{L(L-1)}\\
&= -K \cdot \frac{L-K}{L-1} \cdot Flrn(\psi)(y) \cdot Flrn(\psi)(z).
\end{aligned}$$

3 And:

$$\begin{aligned}
&Cov(pl[K](\psi),\, \pi_y,\, \pi_z)\\
&= (pl[K](\psi) \models \pi_y \;\&\; \pi_z) - (pl[K](\psi) \models \pi_y) \cdot (pl[K](\psi) \models \pi_z)\\
&= K \cdot (K-1) \cdot Flrn(\psi)(y) \cdot \frac{\psi(z)}{L+1} - K \cdot Flrn(\psi)(y) \cdot K \cdot Flrn(\psi)(z)\\
&= K \cdot Flrn(\psi)(y) \cdot \frac{L(K-1) \cdot \psi(z) - K(L+1) \cdot \psi(z)}{L(L+1)}\\
&= -K \cdot \frac{L+K}{L+1} \cdot Flrn(\psi)(y) \cdot Flrn(\psi)(z).
\end{aligned}$$

❑

Definition 5.2.4. For a set X we define the covariances of the three draw dis-

tributions as (joint) multisets over $X \times X$ in:

$$\begin{aligned}
Cov\big(mn[K](\omega)\big) &:= \sum_{y,z\in X} Cov(mn[K](\omega,\pi_y,\pi_z)\,\big|\,y,z\rangle \\
Cov\big(hg[K](\psi)\big) &:= \sum_{y,z\in X} Cov(hg[K](\psi),\pi_y,\pi_z)\,\big|\,y,z\rangle \\
Cov\big(pl[K](\psi)\big) &:= \sum_{y,z\in X} Cov(pl[K](\psi),\pi_y,\pi_z)\,\big|\,y,z\rangle.
\end{aligned}$$

On the diagonal in these joint multisets, where $y = z$, there are the variances. When the elements of the space X are ordered, these covariance multisets can be seen as matrices. We elaborate an illustration.

Example 5.2.5. Consider a group of 50 people of which 25 vote for the green party (G), 10 vote liberal (L) and the ramaining ones vote for the christian-democractic party (C). We thus have a set of vote options $V = \{G, L, C\}$ with a (natural) voter multiset $\nu = 25|G\rangle + 10|L\rangle + 15|C\rangle$.

We select five people from the group and look at their votes. These five people are obtained in hypergeometric mode, where selected individuals step out of the group and are no longer available for subsequent selection.

The hypergeometric mean is introduced in Definition 4.4.1. It gives the multiset:

$$mean\big(hg[5](\nu)\big) \;=\; 5 \cdot Flrn(\nu) \;=\; \tfrac{5}{2}|G\rangle + \tfrac{3}{2}|L\rangle + 1|C\rangle.$$

The deviations of the mean are given by the variance, see Definition 5.2.2.

$$\begin{aligned}
Var\big(hg[5](\nu)\big) \;&=\; 5 \cdot \frac{50-5}{50-1} \cdot \sum_{x\in V} Flrn(\nu)(x) \cdot (1 - Flrn(\nu)(x))\,\big|\,x\rangle \\
&=\; \tfrac{225}{196}|G\rangle + \tfrac{27}{28}|L\rangle + \tfrac{36}{49}|C\rangle \\
&\approx\; 1.148|G\rangle + 0.9643|L\rangle + 0.7347|C\rangle.
\end{aligned}$$

The covariances give a 2-dimensional multiset over $V \times V$, see Definition 5.2.4.

$$\begin{aligned}
Cov\big(hg[5](\nu)\big) \;=\;& \tfrac{225}{196}|G,G\rangle - \tfrac{135}{196}|G,L\rangle - \tfrac{45}{98}|G,C\rangle \\
&- \tfrac{135}{196}|L,G\rangle + \tfrac{27}{28}|L,L\rangle - \tfrac{27}{98}|L,C\rangle \\
&- \tfrac{45}{98}|C,G\rangle - \tfrac{27}{98}|C,L\rangle + \tfrac{36}{49}|C,C\rangle \\
\approx\;& 1.148|G,G\rangle - 0.6888|G,L\rangle - 0.4592|G,C\rangle \\
&- 0.6888|L,G\rangle + 0.9643|L,L\rangle - 0.2755|L,C\rangle \\
&- 0.4592|C,G\rangle - 0.2755|C,L\rangle + 0.7347|C,C\rangle.
\end{aligned}$$

When these covariances are seen as a matrix, we recognise that the matrix is symmetric and has variances on its diagonal.

We recall from Definition 4.4.2 the extension of an observable p on a set X to an observable on natural multisets $\mathcal{N}(X)$ over X. This can be done additively and multiplicatively. Interestingly, the additive extension interacts well with (co)variance in the multinomial case, like with validity in Proposition 4.4.3 (1).

Proposition 5.2.6. *Let $p, q \colon X \to \mathbb{R}$ be observables on a set X, with their addtive extensions $\overline{p}^+, \overline{q}^+ \colon \mathcal{N}(X) \to \mathbb{R}$. For a distribution $\omega \in \mathcal{D}(X)$ and a number $K \in \mathbb{N}$,*

1 *Variance of $\overline{p}^+$ over multinomial draws is related to variance of p over X, via:*

$$\mathit{Var}\Big(\mathit{mn}[K](\omega),\, \overline{p}^+\Big) \;=\; K \cdot \mathit{Var}(\omega, p).$$

2 *Similarly for covariance:*

$$\mathit{Cov}\Big(\mathit{mn}[K](\omega),\, \overline{p}^+,\, \overline{q}^+\Big) \;=\; K \cdot \mathit{Cov}(\omega, p, q).$$

Proof. 1 By Proposition 4.4.3 (1) and Exercise 3.3.7:

$$\begin{aligned}
&\mathit{Var}\Big(\mathit{mn}[K](\omega),\, \overline{p}^+\Big) \\
&= \mathit{mn}[K](\omega) \models \overline{p}^+ \;\&\; \overline{p}^+ - (\mathit{mn}[K](\omega) \models \overline{p}^+)^2 \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \mathit{mn}[K](\omega)(\varphi) \cdot \overline{p}^+(\varphi) \cdot \overline{p}^+(\varphi) - K^2 \cdot (\omega \models p)^2 \\
&= \sum_{x,y \in X} p(x) \cdot p(y) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} \mathit{mn}[K](\omega)(\varphi) \cdot \varphi(x) \cdot \varphi(y) - K^2 \cdot (\omega \models p)^2 \\
&= \sum_{x,y \in X} p(x) \cdot p(y) \cdot \left\{ \begin{array}{ll} K \cdot (K-1) \cdot \omega(x) \cdot \omega(y) & \text{if } x \neq y \\ K \cdot (K-1) \cdot \omega(x)^2 + K \cdot \omega(x) & \text{if } x = y \end{array} \right\} - K^2 \cdot (\omega \models p)^2 \\
&= K \cdot (K-1) \cdot (\omega \models p)^2 + K \cdot (\omega \models p \;\&\; p) - K^2 \cdot (\omega \models p)^2 \\
&= K \cdot (\omega \models p \;\&\; p) - K \cdot (\omega \models p)^2 \\
&= K \cdot \mathit{Var}(\omega, p).
\end{aligned}$$

2 Similarly. ❑

5.2.1 Distributions of validities and of variances

Fix a random variable (ω, p) on a set X and a number K. We can form the multinomial distribution $\mathit{mn}[K](\omega)$ on the set $\mathcal{N}[K](X)$ of natural multisets φ of size K, over X. Applying frequentist learning to such multisets φ gives new distributions $\mathit{Flrn}(\varphi) \in \mathcal{D}(X)$, and thus new random variables $(\mathit{Flrn}(\varphi), p)$. We can look a the validity and variance of the latter. This gives what we call *distributions of validity* and *distributions of variance*. They are defined as distributions

on $\mathbb{R}$ via:

$$\begin{aligned} dval[K](\omega, p) &:= \mathcal{D}(Flrn(-) \models p)\Big(mn[K](\omega)\Big) \\ &= \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi)\, \big|\, Flrn(\varphi) \models p \big\rangle \\ dvar[K](\omega, p) &:= \mathcal{D}(Var(Flrn(-), p))\Big(mn[K](\omega)\Big) \\ &= \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi)\, \big|\, Var(Flrn(\varphi), p) \big\rangle. \end{aligned} \tag{5.1}$$

Such distributions are useful in hypothesis testing in statistics, for instance when ω is a huge distribution for which we wish we check the validity of the observable (or predicate) p. We can then take small samples φ from ω via the multinomial distribution and check the validity of p in the normalised sample $Flrn(\varphi)$. Proposition 5.2.7 (1) below says that the mean of all such samples equals the validity $\omega \models p$.

We illustrate the above definitions (5.1). Consider a three-element set $X = \{a, b, c\}$ with distribution $\omega = \frac{1}{6}|a\rangle + \frac{1}{2}|b\rangle + \frac{1}{3}|c\rangle$ and observable $p = 6 \cdot \mathbf{1}_a + 4 \cdot \mathbf{1}_b + 12 \cdot \mathbf{1}_c$ we get:

$$\begin{aligned} dval[2](\omega, p) &= \tfrac{1}{36}\big|Flrn(2|a\rangle) \models p\big\rangle + \tfrac{1}{6}\big|Flrn(1|a\rangle + 1|b\rangle) \models p\big\rangle \\ &\quad + \tfrac{1}{4}\big|Flrn(2|b\rangle) \models p\big\rangle + \tfrac{1}{9}\big|Flrn(1|a\rangle + 1|c\rangle) \models p\big\rangle \\ &\quad + \tfrac{1}{3}\big|Flrn(1|b\rangle + 1|c\rangle) \models p\big\rangle + \tfrac{1}{9}\big|Flrn(2|c\rangle) \models p\big\rangle \\ &= \tfrac{1}{36}\big|1|a\rangle \models p\big\rangle + \tfrac{1}{6}\big|\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle \models p\big\rangle \\ &\quad + \tfrac{1}{4}\big|1|b\rangle \models p\big\rangle + \tfrac{1}{9}\big|\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|c\rangle \models p\big\rangle \\ &\quad + \tfrac{1}{3}\big|\tfrac{1}{2}|b\rangle + \tfrac{1}{2}|c\rangle \models p\big\rangle + \tfrac{1}{9}\big|1|c\rangle \models p\big\rangle \\ &= \tfrac{1}{36}\big|6\big\rangle + \tfrac{1}{6}\big|5\big\rangle + \tfrac{1}{4}\big|4\big\rangle + \tfrac{1}{9}\big|9\big\rangle + \tfrac{1}{3}\big|8\big\rangle + \tfrac{1}{9}\big|12\big\rangle \end{aligned}$$

Similarly,

$$\begin{aligned} dvar[2](\omega, p) &= \tfrac{1}{36}\big|\, Var(1|a\rangle, p)\big\rangle + \tfrac{1}{6}\big|\, Var(\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle, p)\big\rangle \\ &\quad + \tfrac{1}{4}\big|\, Var(1|b\rangle, p)\big\rangle + \tfrac{1}{9}\big|\, Var(\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|c\rangle, p)\big\rangle \\ &\quad + \tfrac{1}{3}\big|\, Var(\tfrac{1}{2}|b\rangle + \tfrac{1}{2}|c\rangle, p)\big\rangle + \tfrac{1}{9}\big|\, Var(1|c\rangle, p)\big\rangle \\ &= \tfrac{1}{36}\big|6^2 - 6^2\big\rangle + \tfrac{1}{6}\big|26 - 5^2\big\rangle + \tfrac{1}{4}\big|4^2 - 4^2\big\rangle \\ &\quad + \tfrac{1}{9}\big|90 - 9^2\big\rangle + \tfrac{1}{3}\big|80 - 8^2\big\rangle + \tfrac{1}{9}\big|12^2 - 12^2\big\rangle \\ &= \tfrac{7}{18}\big|0\big\rangle + \tfrac{1}{6}\big|1\big\rangle + \tfrac{1}{9}\big|9\big\rangle + \tfrac{1}{3}\big|16\big\rangle. \end{aligned}$$

Diagrammatically the definitions (5.1) involve the composites:

$$\mathcal{D}(X) \xrightarrow{mn[K]} \mathcal{D}(\mathcal{M}[K](X)) \xrightarrow{\mathcal{D}(Flrn)} \mathcal{D}(\mathcal{D}(X)) \overset{\mathcal{D}(-\models p)}{\underset{\mathcal{D}(Var(-,p))}{\rightrightarrows}} \mathcal{D}(\mathbb{R}) \tag{5.2}$$

The distribution of validities $dval[K]$ generalises the *distribution of means* that is often used in hypothesis testing in statistics (see *e.g.* [157]). This distribution of means uses the validiy $(-) \models p$ to compute the mean, by taking the inclusion $incl\colon X \hookrightarrow \mathbb{R}$ as observable, as in Definition 4.1.3. This works of course only when X is a set of numbers. The above approach (5.1) with an arbitrary observable p is more general.

Once we have formed a distribution of validities/variances, we can ask what its validity/variance is. It turns out that these can be expressed in terms of validity/variance of the orginal random variable. These results resemble Theorem 3.3.5 which says that transforming a multinomial distribution along frequentist learning yields the orginal (urn) distribution.

Proposition 5.2.7. *Let (ω, p) be a random variable on a set X, with a number $K > 0$.*

1 *The mean of the distribution of validities is the validity of the original random variable:*

$$mean\Big(dval[K](\omega, p)\Big) = \omega \models p.$$

2 *The variance of the distribution of validities satisfies:*

$$Var\Big(dval[K](\omega, p)\Big) = \frac{Var(\omega, p)}{K}.$$

3 *The mean of the distribution of variances is:*

$$mean\Big(dvar[K](\omega, p)\Big) = \frac{K-1}{K} \cdot Var(\omega, p).$$

There is a fourth option to consider, namely the variance of the distribution of variances, but that doesn't seem to be very interesting.

Proof. 1 Easy, via the following diagrammatic proof.

$$\begin{array}{ccccccc} \mathcal{D}(X) & \xrightarrow{mn[K]} & \mathcal{D}(\mathcal{M}[K](X)) & \xrightarrow{\mathcal{D}(Flrn)} & \mathcal{D}(\mathcal{D}(X)) & \xrightarrow{\mathcal{D}(-\models p)} & \mathcal{D}(\mathbb{R}) \\ & \searrow & & & \downarrow{\scriptstyle flat} & & \downarrow{\scriptstyle mean} \\ & & & & \mathcal{D}(X) & \xrightarrow{(-)\models p} & \mathbb{R} \end{array}$$

The rectangle on the right commutes by Exercise 4.1.8 (3), and the triangle on the left by Theorem 3.3.5.

2 The second item requires more work. Let's assume $supp(\omega) = \{x_1, \ldots, x_n\}$. We first prove the auxiliary result $(*)$ below, via the Multinomial Theorem (1.27) and Exercise 3.3.7, denoted below by the marked equation $\overset{(E)}{=}$.

$$\sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \Big(Flrn(\varphi) \models p\Big)^2 = \frac{(K-1)(\omega \models p)^2 + (\omega \models p^2)}{K}. \quad (*)$$

We reason as follows.

$$
\begin{aligned}
&\sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\Big(Flrn(\varphi)\models p\Big)^2\\
&= \sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\Big(\textstyle\sum_i \frac{\varphi(x_i)}{K}\cdot p(x_i)\Big)^2\\
&\overset{(1.27)}{=} \frac{1}{K^2}\cdot \sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\cdot \sum_{\psi\in\mathcal{N}[2](\{1,\ldots,n\})} \binom{K}{\psi}\cdot \textstyle\prod_i (\varphi(x_i)\cdot p(x_i))^{\psi(i)}\\
&= \frac{1}{K^2}\cdot\Bigg(2\sum_{i\neq j}\sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\cdot\varphi(x_i)\cdot p(x_i)\cdot\varphi(x_j)\cdot p(x_j)\\
&\qquad\qquad + \sum_i \sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\cdot\varphi(x_i)^2\cdot p(x_i)^2\Bigg)\\
&\overset{(E)}{=} \frac{1}{K^2}\cdot\Bigg(2\sum_{i\neq j} K\cdot(K-1)\cdot\omega(x_i)\cdot p(x_i)\cdot\omega(x_j)\cdot p(x_j)\\
&\qquad\qquad + \sum_i K\cdot(K-1)\cdot\omega(x_i)^2\cdot p(x_i)^2 + K\cdot\omega(x_i)\cdot p(x_i)^2\Bigg)\\
&\overset{(1.27)}{=} \frac{K-1}{K}\cdot\big(\textstyle\sum_i \omega(x_i)\cdot p(x_i)\big)^2 + \frac{1}{K}\cdot\big(\textstyle\sum_i \omega(x_i)\cdot p(x_i)^2\big)\\
&= \frac{(K-1)(\omega\models p)^2 + (\omega\models p^2)}{K}.
\end{aligned}
$$

Now we are ready to prove the formula for the variance of the distribution of validities in item (2) in the proposition. We use item (1) and the auxiliarly equation $(*)$.

$$
\begin{aligned}
&Var\Big(dval[K](\omega,p)\Big)\\
&= \sum_{\varphi\in\mathcal{M}[K](X)} mn[K](\omega)(\varphi)\Big(Flrn(\varphi)\models p\Big)^2 - \Big(mean\Big(dval[K](\omega,p)\Big)\Big)^2\\
&\overset{(*)}{=} \frac{(K-1)(\omega\models p)^2 + (\omega\models p^2)}{K} - (\omega\models p)^2\\
&= \frac{-(\omega\models p)^2 + (\omega\models p^2)}{K}\\
&= \frac{Var(\omega,p)}{K}.
\end{aligned}
$$

3 We use item (1) and $(*)$.

$$\begin{aligned}
&\mathit{mean}\Big(\mathit{dvar}[K](\omega, p)\Big) \\
&= \sum_{\varphi \in \mathcal{M}[K](X)} \mathit{mn}[K](\omega)(\varphi) \cdot \mathit{Var}(\mathit{Flrn}(\varphi), p) \\
&= \sum_{\varphi \in \mathcal{M}[K](X)} \mathit{mn}[K](\omega)(\varphi) \cdot \Big((\mathit{Flrn}(\varphi) \models p^2) - (\mathit{Flrn}(\varphi) \models p)^2\Big) \\
&\overset{(*)}{=} (\omega \models p^2) - \frac{(K-1)(\omega \models p)^2 + (\omega \models p^2)}{K} \\
&= \frac{(K-1)(\omega \models p^2)}{K} - \frac{(K-1)(\omega \models p)^2}{K} \\
&= \frac{K-1}{K} \cdot \mathit{Var}(\omega, p).
\end{aligned}$$

❑

Exercises

5.2.1 In the distribution of validities (5.1) we have used multinomial samples. We can also use hypergeometric ones. Show that in that case the analogue of Proposition 5.2.7 (1) still holds: for an urn $\upsilon \in \mathcal{N}(X)$, an observable $p\colon X \to \mathbb{R}$ and a number $K \leq \|\upsilon\|$,

$$\mathit{mean}\Big(\mathcal{D}(\mathit{Flrn}(-) \models p)(\mathit{hg}[K](\upsilon))\Big) = \mathit{Flrn}(\upsilon) \models p.$$

5.3 Joint-state covariance and correlation

Section 5.1 has introduced variance and *shared-state* covariance / correlation. This section looks at a slightly different version, which we call *joint-state* covariance / correlation. It is subtly different from the shared-state version since the observables involved are defined not on the sample space of the shared underlying state, but on the components of a product space.

We will thus first define covariance and correlation for a joint state with observables on its component spaces, as a special case of what we have seen so far.

Definition 5.3.1. Let $\tau \in \mathcal{D}(X_1 \times X_2)$ be a joint state on sets X_1, X_2 and let $q_1 \in \mathit{Obs}(X_1)$ and $q_2 \in \mathit{Obs}(X_2)$ be two observables on these two sets *separately*. In this situation the *joint covariance* is defined as:

$$\mathit{JCov}(\tau, q_1, q_2) \coloneqq \mathit{Cov}(\tau,\, \pi_1 \mathrel{=\!\!\ll} q_1,\, \pi_2 \mathrel{=\!\!\ll} q_2).$$

Thus we use the weakenings $\pi_1 \mathrel{=\!\!\ll} q_1 = q_1 \otimes \mathbf{1}$ and $\pi_2 \mathrel{=\!\!\ll} q_2 = \mathbf{1} \otimes q_2$ to turn

the observables q_1, q_2 on different sets X_1, X_2 into observables on the same (product) set $X_1 \times X_2$ — so that Definition 5.1.4 applies.

Similarly, the *joint correlation* is:

$$JCor(\tau, q_1, q_2) \coloneqq Cor(\tau, \pi_1 \mathrel{=\!\!\ll} q_1, \pi_2 \mathrel{=\!\!\ll} q_2).$$

In both these cases, if there are inclusions $X_1 \hookrightarrow \mathbb{R}$ and $X_2 \hookrightarrow \mathbb{R}$, then one can use these inclusions as random variables and write just $JCov(\tau)$ and $JCor(\tau)$.

Joint covariance can be reformulated in different ways, including in the style that we have seen before, in Lemma 5.1.3 and 5.1.5.

Lemma 5.3.2. *Joint covariance can be reformulated as:*

$$\begin{aligned} JCov(\tau, q_1, q_2) &= \tau \models (q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1}) \otimes (q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1}). \\ &= (\tau \models q_1 \otimes q_2) - (\tau[1,0] \models q_1) \cdot (\tau[0,1] \models q_2). \end{aligned}$$

We see that if the joint state τ is non-entwined, that is, if it is the product of its marginals $\tau[1,0]$ and $\tau[0,1]$, then the joint covariance is 0, whatever the obervables q_1, q_2 are. This situation will be investigated further in the next section.

Proof. The first equation follows from:

$$\begin{aligned} JCov(\tau, q_1, q_2) &= Cov(\tau, \pi_1 \mathrel{=\!\!\ll} q_1, \pi_2 \mathrel{=\!\!\ll} q_2). \\ &= \tau \models ((\pi_1 \mathrel{=\!\!\ll} q_1) - (\tau \models \pi_1 \mathrel{=\!\!\ll} q_1) \cdot \mathbf{1}) \\ &\qquad \& \ ((\pi_2 \mathrel{=\!\!\ll} q_2) - (\tau \models \pi_2 \mathrel{=\!\!\ll} q_2) \cdot \mathbf{1}) \\ &= \tau \models ((\pi_1 \mathrel{=\!\!\ll} q_1) - (\tau[1,0] \models q_1) \cdot (\pi_1 \mathrel{=\!\!\ll} \mathbf{1})) \\ &\qquad \& \ ((\pi_2 \mathrel{=\!\!\ll} q_2) - (\tau[0,1] \models q_2) \cdot (\pi_2 \mathrel{=\!\!\ll} \mathbf{1})) \\ &= \tau \models \Big(\pi_1 \mathrel{=\!\!\ll} (q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1})\Big) \\ &\qquad \& \ \Big(\pi_2 \mathrel{=\!\!\ll} (q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1})\Big) \\ &= \tau \models (q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1}) \otimes (q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1}). \end{aligned}$$

The last equation follows from Exercise 4.3.7.

Via this first equation in Lemma 5.3.2 we prove the second one:

$$\begin{aligned} &JCov(\tau, q_1, q_2) \\ &= \tau \models (q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1}) \otimes (q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1}) \\ &= \textstyle\sum_{x,y} \tau(x,y) \cdot (q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1})(x) \cdot (q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1})(y) \\ &= \textstyle\sum_{x,y} \tau(x,y) \cdot (q_1(x) - (\tau[1,0] \models q_1)) \cdot (q_2(y) - (\tau[0,1] \models q_2)) \\ &= \big(\textstyle\sum_{x,y} \tau(x,y) \cdot q_1(x) \cdot q_2(y)\big) - \big(\textstyle\sum_{x,y} \tau(x,y) \cdot q_1(x) \cdot (\tau[0,1] \models q_2)\big) \\ &\qquad - \big(\textstyle\sum_{x,y} \tau(x,y) \cdot (\tau[1,0] \models q_1) \cdot q_2(y)\big) + (\tau[1,0] \models q_1)(\tau[0,1] \models q_2) \\ &= (\tau \models q_1 \otimes q_2) - (\tau[1,0] \models q_1)(\tau[0,1] \models q_2). \end{aligned}$$

❑

We turn to some illustrations. Many examples of covariance are actually of joint form, especially if the underlying sets are subsets of the real numbers. In the joint case it makes sense to leave these inclusions implicit, as will be illustrated below.

Example 5.3.3. 1 Consider sets $X = \{1, 2\}$ and $Y = \{1, 2, 3\}$ as subsets of $\mathbb{R}$, together with a joint distribution $\tau \in \mathcal{D}(X \times Y)$ given by:

$$\tau = \tfrac{1}{4}|1, 1\rangle + \tfrac{1}{4}|1, 2\rangle + \tfrac{1}{4}|2, 2\rangle + \tfrac{1}{4}|2, 3\rangle.$$

Its two marginals are:

$$\tau[1, 0] = \tfrac{1}{2}|1\rangle + \tfrac{1}{2}|2\rangle \qquad \tau[0, 1] = \tfrac{1}{4}|1\rangle + \tfrac{1}{2}|2\rangle + \tfrac{1}{4}|3\rangle.$$

Since both $X \hookrightarrow \mathbb{R}$ and $Y \hookrightarrow \mathbb{R}$ we get means as validities of the inclusions:

$$\mathit{mean}(\tau[1, 0]) = \tfrac{3}{2} \qquad \mathit{mean}(\tau[1, 0]) = 2.$$

Now we can compute the joint covariance in the joint state τ as:

$$\begin{aligned} \mathit{JCov}(\tau) &= \tau \models (\mathit{incl} - \mathit{mean}(\tau[1, 0]) \cdot \mathbf{1}) \otimes (\mathit{incl} - \mathit{mean}(\tau[1, 0]) \cdot \mathbf{1}) \\ &= \textstyle\sum_{x,y} \tau(x, y) \cdot (x - \frac{3}{2}) \cdot (y - 2) \\ &= \tfrac{1}{4}(-\tfrac{1}{2} \cdot -1 + \tfrac{1}{2} \cdot 1) \\ &= \tfrac{1}{4}. \end{aligned}$$

2 In order to calculate the (joint) correlation of τ we first need the variances of its marginals:

$$\begin{aligned} \mathit{Var}(\tau[1, 0]) &= \textstyle\sum_x \tau[1, 0](x) \cdot (x - \frac{3}{2})^2 = \tfrac{1}{4} \\ \mathit{Var}(\tau[0, 1]) &= \textstyle\sum_y \tau[0, 1](y) \cdot (y - 2)^2 = \tfrac{1}{2}. \end{aligned}$$

Then:

$$\mathit{JCor}(\tau) = \frac{\mathit{JCov}(\tau)}{\sqrt{\mathit{Var}(\tau[1, 0])} \cdot \sqrt{\mathit{Var}(\tau[0, 1])}} = \frac{1/4}{1/2 \cdot 1/\sqrt{2}} = \tfrac{1}{2}\sqrt{2}.$$

We have defined joint covariance as a special case of ordinary covariance. We now show that shared-state covariance can also be seen as joint-state covariance, namely for a copied state. Recall that copying of states is a subtle matter, since $\Delta \gg\!= \omega \neq \omega \otimes \omega$, in general, see Subsection 2.5.2.

Proposition 5.3.4. *Shared state covariance can be expressed as joint covariance:*

$$\mathit{Cov}(\omega, p_1, p_2) = \mathit{JCov}(\Delta \gg\!= \omega, p_1, p_2).$$

More generally, for suitably typed channels c, d,

$$\mathit{Cov}(\omega,\ c \mathrel{=\!\!\ll} p,\ d \mathrel{=\!\!\ll} q) = \mathit{JCov}(\langle c, d\rangle \gg\!= \omega,\ p,\ q).$$

Proof. We prove the second equation since the first one is a special case, namely when c, d are identity channels. Via Lemma 5.3.2 we get:

$$\begin{aligned}
&JCov(\langle c, d\rangle \gg\!= \omega, p, q) \\
&= \langle c, d\rangle \gg\!= \omega \models (p - ((\langle c, d\rangle \gg\!= \omega)[1, 0] \models p) \cdot \mathbf{1}) \\
&\qquad\qquad \otimes (q - ((\langle c, d\rangle \gg\!= \omega)[0, 1] \models q) \cdot \mathbf{1}) \\
&= \omega \models \langle c, d\rangle \mathrel{=\!\!\ll} \Big((p - (c \gg\!= \omega \models p) \cdot \mathbf{1}) \otimes (q - (d \gg\!= \omega \models q) \cdot \mathbf{1})\Big) \\
&= \omega \models \Big(c \mathrel{=\!\!\ll} (p - (\omega \models c \mathrel{=\!\!\ll} p) \cdot \mathbf{1})\Big) \mathbin{\&} \Big(d \mathrel{=\!\!\ll} (q - (\omega \models d \mathrel{=\!\!\ll} q) \cdot \mathbf{1})\Big) \\
&\qquad \text{by Exercise 4.3.8} \\
&= \omega \models ((c \mathrel{=\!\!\ll} p) - (\omega \models c \mathrel{=\!\!\ll} p) \cdot \mathbf{1}) \mathbin{\&} ((d \mathrel{=\!\!\ll} q) - (\omega \models d \mathrel{=\!\!\ll} q) \cdot \mathbf{1}) \\
&\qquad \text{since } c \mathrel{=\!\!\ll} (-) \text{ is linear and preserves } \mathbf{1} \\
&= Cov(\omega, c \mathrel{=\!\!\ll} p, d \mathrel{=\!\!\ll} q).
\end{aligned}$$

❑

We started with 'ordinary' covariance in Definition 5.1.4 for two random variables with a shared state. It was subsequently used to define the 'joint' version in Definition 5.3.1. The above result shows that we could have done this the other way around: obtain the ordinary formulation in terms of the joint version. As we shall see below, there are notable differences between shared-state and joint-state versions, see Proposition 5.4.3 and Theorem 5.4.6 in the next section.

But first we formulate a joint-state analogue for the linearity properties of Theorem 5.1.7.

Theorem 5.3.5. *Consider a state $\tau \in \mathcal{D}(X_1 \times X_2)$, with observables $q_1 \in Obs(X_1)$, $q_2, q_3 \in Obs(X_2)$ and numbers $r, s \in \mathbb{R}$.*

1 Joint-state covariance satisfies:

$$\begin{aligned}
JCov(\tau, q_1, q_2) &= JCov(\tau, q_2, q_1) \\
JCov(\tau, q_1, \mathbf{1}) &= 0 \\
JCov(\tau, r \cdot q_1, q_2) &= r \cdot JCov(\tau, q_1, q_2) \\
JCov(\omega, q_1, q_2 + q_3) &= JCov(\tau, q_1, q_2) + Cov(\tau, q_1, q_3). \\
JCov(\tau, q_1 + r \cdot \mathbf{1}, q_2 + s \cdot \mathbf{1}) &= JCov(\tau, q_1, q_2).
\end{aligned}$$

2 For joint-state correlation we have:

$$\begin{aligned}
JCor(\tau, r \cdot q_1, s \cdot q_2) &= \begin{cases} JCor(\tau, q_1, q_2) & \text{if } r, s \text{ have the same sign} \\ -JCor(\tau, q_1, q_2) & \text{otherwise.} \end{cases} \\
JCor(\tau, q_1 + r \cdot \mathbf{1}, q_2 + s \cdot \mathbf{1}) &= JCor(\omega, q_1, q_2).
\end{aligned}$$

Proof. This follows directly from Theorem 5.1.7, using that predicate trans-

formation $\pi_i \mathrel{=\!\!\ll} (-)$ is linear and thus preserves sums and scalar multiplications (and also truth), see Lemma 4.3.2 (2). ❑

We conclude this section with a medical example about the correlation between disease and test.

Example 5.3.6. We start with a space $D = \{d, d^\perp\}$ for occurence of a disease or not (for a particular person) and a space $T = \{p, n\}$ for a positive or negative test outcome. Prevalence is used to indicate the prior likelihood of occurrence of the disease, for instance in the whole population, before a test. It can be described via a flip-like channel:

$$[0,1] \xrightarrow{\ prev\ } D \quad \text{with} \quad prev(r) \coloneqq r|d\rangle + (1-r)|d^\perp\rangle.$$

We assume that there is a test for the disease with the following characteristics.

- ('sensitivity') If someone has the disease, then the test is positive with probability of 90%.
- ('specificity') If someone does not have the disease, there is a 95% chance that the test is negative.

We formalise this via a channel $test\colon D \nrightarrow T$ with:

$$test(d) = \tfrac{9}{10}|p\rangle + \tfrac{1}{10}|n\rangle \qquad test(d^\perp) = \tfrac{1}{20}|p\rangle + \tfrac{19}{20}|n\rangle.$$

We can now form the joint 'graph' state:

$$joint(r) \coloneqq \langle id, test\rangle \gg\!= prev(r) \in \mathcal{D}(D \times T).$$

Exercise 5.3.3 below tells that it does not really matter which observables we choose, so we simply take $\mathbf{1}_d\colon D \to [0,1]$ and $\mathbf{1}_p\colon T \to [0,1]$, mapping d and p to 1 and $d^\perp$ and n to 0. We are thus interested in (joint-state) correlation function:

$$[0,1] \ni r \longmapsto JCor\Big(joint(r), \mathbf{1}_d, \mathbf{1}_p\Big) \in [-1,1].$$

This is plotted in Figure 5.1, on the left. We see that, with the sensitivity and specificity values as given above, there is a clear positive correlation beteen disease and test, but less so in the corner cases with minimal and maximal prevalence.

We now fix a prevalence of 20% and wish to understand correlation as a function of sensitivity and specificity. We thus parameterise the above test channel to $test(se, sp)\colon D \nrightarrow T$ with parameters $se, sp \in [0,1]$.

$$test(se,sp)(d) = se|p\rangle + (1-se)|n\rangle$$
$$test(se,sp)(d^\perp) = (1-sp)|p\rangle + sp|n\rangle.$$

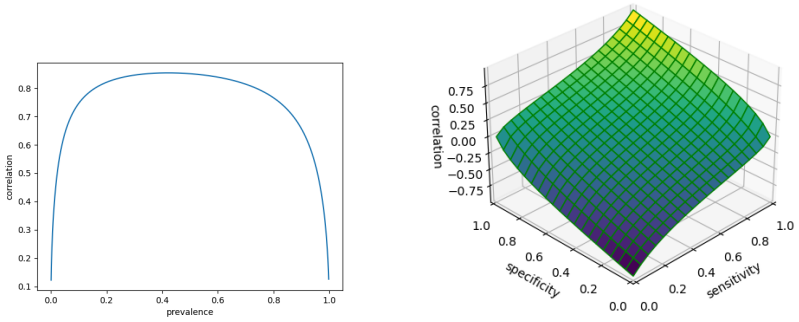


Figure 5.1 Disease-Test correlations, on the left as a function of prevalence (with fixed sensitivity and specificity) and on the right as a function of sensitivity and specificity (with fixed prevalence), see Example 5.3.6 for details.

As before we form a joint state, but now with different parameters.

$$joint(se, sp) \coloneqq \langle id, test(se, sp)\rangle \gg\!= prev(\tfrac{1}{5}) \in \mathcal{D}(D \times T).$$

We are then interested in the function:

$$[0,1] \times [0,1] \ni (se, sp) \longmapsto JCor\Big(joint(se, sp), \mathbf{1}_d, \mathbf{1}_p\Big) \in [-1, 1].$$

It is described on the right in Figure 5.1. We see that with maximal sensitivity and specificity (both 1) the correlation between disease and test is also maximal (actually 1), and dually with minimal sensitivity and specificity (both 0) the correlation is minimal (namely -1). These (unrealistic) extremes correspond to an optimal test and an inferior one.

Exercise 5.3.4 makes some intuitive properties of this (parameterised) correlation between disease and test explicit.

Exercises

5.3.1 Find examples of covariance and correlation computations in the literature (or online) and determine if they are of shared-state or joint-state form.

5.3.2 Prove that:

$$JCor(\tau, q_1, q_2) = \frac{JCov(\tau, q_1, q_2)}{StDev(\tau[1,0], q_1) \cdot StDev(\tau[0,1], q_2)}.$$

5.3.3 Consider two two-element sample spaces $X_1 = \{a_1, b_1\}$ and $X_2 = \{a_2, b_2\}$ with a joint state $\tau \in \mathcal{D}(X_1 \times X_2)$. Prove that in this binary case joint-state covariance and correlation do not depend on the observables, that is:

$$JCor(\tau, p_1, p_2) = \pm JCor(\tau, q_1, q_2),$$

for all *non-constant* observables $p_1, q_1 \in Obs(X_1)$, $p_2, q_2 \in Obs(X_2)$. *Hint:* Use Theorem 5.3.5 to massage p_1, p_2 to the observables which send a_1, b_1 to 0 and a_2, b_2 to 1.

5.3.4 Prove in the context of Example 5.3.6,

1 for each $s \in [0, 1]$ one has:

$$JCor\Big(joint(s, 1-s, \mathbf{1}_d, \mathbf{1}_p\Big) = 0;$$

2 for all $se, sp \in [0, 1]$,

$$se + sp \geq 1 \iff JCor\Big(joint(se, sp), \mathbf{1}_d, \mathbf{1}_p\Big) \geq 0.$$

5.4 Independence for random variables

Earlier we have called a joint state entwined when it cannot be written as product of its marginals. This is sometimes called dependence. But we like to reserve the word dependence for random variables in this setting. Such dependence (or independence) is the topic of this section. We shall follow the approach of the previous two sections and introduce two versions of dependence, also called *shared-state* and *joint-state*. Independence is related to the property 'covariance is zero', but in a subtle manner. This will be elaborated below.

Suppose we have two random variables describing the number of icecream sales and the temperature. Intuitively one expects a dependence between the two, and that the two variables are correlated (in an informal sense). The opposite, namely *independence* is usually formalised as follows. Two random variables p_1, p_2 are called independent if the equation,

$$P[p_1 = a, p_2 = b] = P[p_1 = a] \cdot P[p_2 = b]. \tag{5.3}$$

holds for all real numbers a, b. In this formulation a distribution is assumed in the background. We like to use it explicitly. How should the above equation (5.3) then be read?

As we described in Subsection 4.1.1, the expression $P[p = a]$ can be interpreted as transformation along the observable p_i, considered as deterministic channel:

$$P[p = a] \;=\; p \gg\!= \omega \models \mathbf{1}_a \overset{(4.9)}{=} \omega \models p \ll\!= \mathbf{1}_a.$$

where ω is the implicit distribution.

We can then translate the above equation (5.3) into:

$$\langle p_1, p_2 \rangle \gg\!= \omega \;=\; (p_1 \gg\!= \omega) \otimes (p_2 \gg\!= \omega). \tag{5.4}$$

But this says that the joint state $\langle p_1, p_2 \rangle \gg\!= \omega$ on $\mathbb{R} \times \mathbb{R}$, transformed along $\langle p_1, p_2 \rangle \colon X \to \mathbb{R} \times \mathbb{R}$, is non-entwined: it is the product of its marginals. Indeed, its (first) marginal is:

$$\begin{aligned} (\langle p_1, p_2 \rangle \gg\!= \omega)[1, 0] &= \pi_1 \gg\!= (\langle p_1, p_2 \rangle \gg\!= \omega) \\ &= (\pi_1 \circ \langle p_1, p_2 \rangle) \gg\!= \omega \\ &= p_1 \gg\!= \omega. \end{aligned}$$

This brings us to the following formulation. We formulate it for two random variables, but it can easily be extended to n-ary form.

Definition 5.4.1. Let (ω, p_1) and (ω, p_2) be two random variables with a common, shared state ω. They will be called *independent* if the transformed state $\langle p_1, p_2 \rangle \gg\!= \omega$ on $\mathbb{R} \times \mathbb{R}$ is non-entwined, as in Equation (5.4): it is the product of its marginals:

$$\langle p_1, p_2 \rangle \gg\!= \omega \;=\; (p_1 \gg\!= \omega) \otimes (p_2 \gg\!= \omega).$$

We sometimes call this the *shared-state* form of independence, in order to distinguish it from a later *joint-state* version.

We give an illustration, of non-independence, that is, of dependence.

Example 5.4.2. Consider a fair coin $\mathit{flip} = \frac{1}{2}|1\rangle + \frac{1}{2}|0\rangle$. We are going to use it twice, first to determine how much we will bet (either €100 or €50), and secondly to determine whether the bet is won or not. Thus we use the distribution $\omega = \mathit{flip} \otimes \mathit{flip}$ with underlying set 2×2, where $2 = \{0, 1\}$. We will define two observables $p_1, p_2 \colon 2 \times 2 \to \mathbb{R}$, to be used as random variables for this same distribution ω.

We first define an auxiliary observable $p \colon 2 \to \mathbb{R}$ for the amount of the bet:

$$p(1) \;=\; 100 \qquad\qquad p(0) \;=\; 50.$$

We then define $p_1 = p \otimes \mathbf{1} = p \circ \pi_1 \colon 2 \times 2 \to \mathbb{R}$, as on the left below. The

observable p_2 is defined on the right.

$$p_1(x,y) \coloneqq \begin{cases} 100 & \text{if } x = 1 \\ 50 & \text{if } x = 0 \end{cases} \qquad p_2(x,y) \coloneqq \begin{cases} p(x) & \text{if } y = 1 \\ -p(x) & \text{if } y = 0. \end{cases}$$

We claim that (ω, p_1) and (ω, p_2) are *not* independent. Intuitively this may be clear, since the observable p forms a connection between p_1 and p_2. Formally, we can see this by doing the calculations:

$$\begin{aligned} \langle p_1, p_2 \rangle \gg\!= \omega &= \mathcal{D}(\langle p_1, p_2 \rangle)(\mathit{flip} \otimes \mathit{flip}) \\ &= \tfrac{1}{4}\big|p_1(1,1), p_2(1,1)\big\rangle + \tfrac{1}{4}\big|p_1(1,0), p_2(1,0)\big\rangle \\ &\qquad + \tfrac{1}{4}\big|p_1(0,1), p_2(0,1)\big\rangle + \tfrac{1}{4}\big|p_1(0,0), p_2(0,0)\big\rangle \\ &= \tfrac{1}{4}\big|100, 100\big\rangle + \tfrac{1}{4}\big|100, -100\big\rangle + \tfrac{1}{4}\big|50, 50\big\rangle + \tfrac{1}{4}\big|50, -50\big\rangle. \end{aligned}$$

The two marginals are:

$$\begin{aligned} p_1 \gg\!= \omega &= (\langle p_1, p_2 \rangle \gg\!= \omega)[1,0] = \tfrac{1}{2}|100\rangle + \tfrac{1}{2}|50\rangle \\ p_2 \gg\!= \omega &= (\langle p_1, p_2 \rangle \gg\!= \omega)[0,1] = \tfrac{1}{4}|100\rangle + \tfrac{1}{4}|-100\rangle + \tfrac{1}{4}|50\rangle + \tfrac{1}{4}|-50\rangle. \end{aligned}$$

It is not hard to see that the parallel product $\otimes$ of these two marginal distributions differs from $\langle p_1, p_2 \rangle \gg\!= \omega$.

Proposition 5.4.3. *The shared-state covariance of shared-state independent random variables is zero: if random variables (ω, p_1) and (ω, p_2) are independent, then $\mathit{Cov}(\omega, p_1, p_2) = 0$.*

The converse does not hold.

Proof. If (ω, p_1) and (ω, p_2) are independent, then, by definition, $\langle p_1, p_2 \rangle \gg\!= \omega = (p_1 \gg\!= \omega) \otimes (p_2 \gg\!= \omega)$. The calculation belows shows that covariance is then zero. It uses multiplication $\&\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as observable. It can also be described as the parallel product $\mathit{id} \otimes \mathit{id}$ of the observable $\mathit{id}\colon \mathbb{R} \to \mathbb{R}$ with itself.

$$\begin{aligned} &\mathit{Cov}(\omega, p_1, p_2) \\ &= (\omega \models p_1 \mathbin{\&} p_2) - (\omega \models p_1) \cdot (\omega \models p_2) && \text{by Lemma 5.1.5} \\ &= (\omega \models \& \circ \langle p_1, p_2 \rangle) \\ &\qquad - (p_1 \gg\!= \omega \models \mathit{id}) \cdot (p_2 \gg\!= \omega \models \mathit{id}) && \text{by Exercise 4.1.4} \\ &= (\langle p_1, p_2 \rangle \gg\!= \omega \models \&) \\ &\qquad - ((p_1 \gg\!= \omega) \otimes (p_2 \gg\!= \omega) \models \mathit{id} \otimes \mathit{id}) && \text{by Lemma 4.2.8, 4.1.4} \\ &= (\langle p_1, p_2 \rangle \gg\!= \omega \models \&) - (\langle p_1, p_2 \rangle \gg\!= \omega \models \&) && \text{by assumption} \\ &= 0. \end{aligned}$$

The claim that the converse does not hold follows from Example 5.4.4, right below. ❑

Example 5.4.4. We continue Example 5.4.2. The set-up used there involves two dependent random variables (ω, p_1) and (ω, p_2), with shared state $\omega = \mathit{flip} \otimes \mathit{flip}$. We show here that they (nevertheless) have covariance zero. This proves the second claim of Proposition 5.4.3, namely that zero-covariance need not imply independence, in the shared-state context.

We first compute the validities (expected values):

$$\begin{array}{rcl} \omega \models p_1 & = & \frac{1}{4} \cdot p_1(1,1) + \frac{1}{4} \cdot p_1(1,0) + \frac{1}{4} \cdot p_1(0,1) + \frac{1}{4} \cdot p_1(0,0) \\ & = & \frac{1}{4} \cdot 100 + \frac{1}{4} \cdot 100 + \frac{1}{4} \cdot 50 + \frac{1}{4} \cdot 50 \;=\; 75 \\ \omega \models p_2 & = & \frac{1}{4} \cdot 100 + \frac{1}{4} \cdot -100 + \frac{1}{4} \cdot 50 + \frac{1}{4} \cdot -50 \;=\; 0 \end{array}$$

Then:

$$\begin{array}{l} \mathit{Cov}(\omega, p_1, p_2) \\ \quad = \omega \models (p_1 - (\omega \models p_1) \cdot \mathbf{1}) \mathbin{\&} (p_2 - (\omega \models p_2) \cdot \mathbf{1}) \\ \quad = \frac{1}{4}\Big((100-75) \cdot 100 + (100-75) \cdot -100 + (50-75) \cdot 50 + (50-75) \cdot -50\Big) \\ \quad = 0. \end{array}$$

We now turn to independence in *joint-state* form, in analogy with joint-state covariance in Definition 5.3.1.

Definition 5.4.5. Let $\tau \in \mathcal{D}(X_1 \times X_2)$ be a joint state and with two observables $q_1 \in \mathit{Obs}(X_1)$ and $q_2 \in \mathit{Obs}(X_2)$. We say that there is *joint-state* independence of q_1, q_2 if the two random variables $(\tau, \pi_1 \mathrel{=\!\!\ll} q_1)$ and $(\tau, \pi_2 \mathrel{=\!\!\ll} q_2)$ are (shared-state) independent, as described in Definition 5.4.1.

Concretely, this means that:

$$(q_1 \otimes q_2) \gg\!= \tau \;=\; (q_1 \gg\!= \tau[1,0]) \otimes (q_2 \gg\!= \tau[0,1]). \tag{5.5}$$

Equation (5.5) is an instance of the formulation (5.4) used in Definition 5.4.1 since $\pi_i \mathrel{=\!\!\ll} q_i = q_i \circ \pi_i$ and:

$$\begin{array}{rcl} \langle q_1 \circ \pi_1,\, q_2 \circ \pi_2 \rangle \gg\!= \tau & = & \mathcal{D}(\langle q_1 \circ \pi_1,\, q_2 \circ \pi_2 \rangle)(\tau) \\ & = & \mathcal{D}(q_1 \times q_2)(\tau) \\ & = & (q_1 \otimes q_2) \gg\!= \tau \\ ((q_1 \circ \pi_1) \gg\!= \tau) \otimes ((q_2 \circ \pi_2) \gg\!= \tau) & = & (q_1 \gg\!= (\pi_1 \gg\!= \tau)) \otimes (q_2 \gg\!= (\pi_2 \gg\!= \tau)) \\ & = & (q_1 \gg\!= \tau[1,0]) \otimes (q_2 \gg\!= \tau[0,1]). \end{array}$$

In the joint-state case — unlike in the shared-state situation — there is a tight connection between non-entwinedness, independence and covariance being zero.

Theorem 5.4.6. *For a joint state τ the following three statements are equivalent.*

1 τ is non-entwined, i.e. is the product of its marginals;
2 all observables $q_i \in Obs(X_i)$ are joint-state independent wrt. τ;
3 the joint-state covariance $JCov(\tau, q_1, q_2)$ is zero, for all observables $q_i \in Obs(X_i)$ — or equivalently, all correlations $JCor(\tau, q_1, q_2)$ are zero.

Proof. Let joint state $\tau \in \mathcal{D}(X_1 \times X_2)$ be given. We write $\tau_i := \pi_i \gg\!= \tau \in \mathcal{D}(X_i)$ for its marginals.

(1) $\Rightarrow$ (2). If τ is non-entwined, then $\tau = \tau_1 \otimes \tau_2$. Hence for all observables $q_1 \in Obs(X_1)$ and $q_2 \in Obs(X_2)$ we have that $\sigma := (q_1 \otimes q_2) \gg\!= \tau$ is non-entwined. To see this, first note that $\pi_i \gg\!= \sigma = q_i \gg\!= \tau_i$. Then, by Lemma 2.4.2 (1):

$$\begin{aligned}(\pi_1 \gg\!= \sigma) \otimes (\pi_2 \gg\!= \sigma) &= (q_1 \gg\!= \tau_1) \otimes (q_2 \gg\!= \tau_2) \\ &= (q_1 \otimes q_2) \gg\!= (\tau_1 \otimes \tau_2) = (q_1 \otimes q_2) \gg\!= \tau = \sigma.\end{aligned}$$

(2) $\Rightarrow$ (3). Let $q_1 \in Obs(X_1)$ and $q_2 \in Obs(X_2)$ be two observables. We may assume that q_1, q_2 are independent wrt. τ, that is, $(q_1 \otimes q_2) \gg\!= \tau = (q_1 \gg\!= \tau_1) \otimes (q_2 \gg\!= \tau_2)$ as in (5.5). We must prove $JCov(\tau, q_1, q_2) = 0$. Consider, like in the proof of Proposition 5.4.3, the multiplication map $\&: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, given by $\&(r, s) = r \cdot s$, as an observable on $\mathbb{R} \times \mathbb{R}$. We consider its validity:

$$\begin{aligned}(q_1 \otimes q_2) \gg\!= \tau \models \& &= \textstyle\sum_{r,s} ((q_1 \otimes q_2) \gg\!= \tau)(r, s) \cdot \&(r, s) \\ &= \textstyle\sum_{r,s} \mathcal{D}(q_1 \times q_2)(\tau)(r, s) \cdot r \cdot s \\ &= \textstyle\sum_{r,s} \big(\sum_{(x,y) \in (q_1 \times q_2)^{-1}(r,s)} \tau(x, y) \big) \cdot r \cdot s \\ &= \textstyle\sum_{r,s} \sum_{x \in q_1^{-1}(r), y \in q_2^{-1}(s)} \tau(x, y) \cdot q_1(x) \cdot q_2(y) \\ &= \textstyle\sum_{x,y} \tau(x, y) \cdot (q_1 \otimes q_2)(x, y) \\ &= \tau \models q_1 \otimes q_2.\end{aligned}$$

In the same way one proves $(q_1 \gg\!= \tau_1) \otimes (q_2 \gg\!= \tau_2) \models \& = (\tau_1 \models q_1) \cdot (\tau_2 \models q_2)$. But then we are done via the formulation of binary covariance from Lemma 5.3.2:

$$\begin{aligned}JCov(\tau, q_1, q_2) &= (\tau \models q_1 \otimes q_2) - (\tau_1 \models q_1) \cdot (\tau_2 \models q_2) \\ &= ((q_1 \otimes q_2) \gg\!= \tau \models \&) - ((q_1 \gg\!= \tau_1) \otimes (q_2 \gg\!= \tau_2) \models \&) \\ &= ((q_1 \otimes q_2) \gg\!= \tau \models \&) - ((q_1 \otimes q_2) \gg\!= \tau \models \&) \\ &= 0.\end{aligned}$$

(3) $\Rightarrow$ (1). Let joint-state covariance $JCov(\tau, q_1, q_2)$ be zero for all observables q_1, q_2. In order to prove that τ is non-entwined, we have to show $\tau(x, y) = \tau_1(x) \cdot \tau_2(y)$ for all $(x, y) \in X_1 \times X_2$. We choose as random variables the observables $\mathbf{1}_x$ and $\mathbf{1}_y$ and use again the formulation of covariance from Lemma 5.3.2. Then, since, by assumption, the binary covariance is zero, so that:

$$\tau(x, y) = \tau \models \mathbf{1}_x \otimes \mathbf{1}_y = (\tau_1 \models \mathbf{1}_x) \cdot (\tau_2 \models \mathbf{1}_y) = \tau_1(x) \cdot \tau_2(y). \qquad \square$$

In essence this result says that joint-state independence and joint-state covariance being zero are not properties of observables, but of joint states.

Exercises

5.4.1 Prove, in the setting of Definition 5.4.5 that the first marginal $((q_1 \otimes q_2) \gg\!= \tau)[1, 0]$ of the transformed state along $q_1 \otimes q_2$ is equal to the transformed marginal $q_1 \gg\!= \tau[1, 0]$.

5.5 The law of large numbers, in weak form

In this section we describe what is called the weak law of large numbers. The strong version appears later on, in **??**. This weak law captures limit behaviour of probabilistic operations, such as: if we throw a fair dice many, many times, we expect to see each number of pips $\frac{1}{6}$ of the time. It is sometimes also called Bernoulli's Theorem.

We shall describe this law of large numbers in two forms: as binary version, which is most familiar, and as multivariate version. Both versions use results about means and variances that we have seen before.

First we need to introduced two famous inequalities, due to Markov and to Chebyshev. If we have an observable $p\colon X \to \mathbb{R}$ and a number $a \in \mathbb{R}$ we introduce a sharp predicate $p \geq a$ on X, defined in an obvious way as:

$$(p \geq a)(x) \coloneqq \begin{cases} 1 & \text{if } p(x) \geq a \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

We can similarly write $p > a$, $p \leq a$ and $p < a$.

Lemma 5.5.1. *Let $\omega \in \mathcal{D}(X)$ be a state, $p \in \mathit{Obs}(X)$ an observable, and $a \in \mathbb{R}$ an arbitrary number. Then:*

1 *Chebyshev's inequality holds: $a \cdot \big(\omega \models (p \geq a)\big) \leq \omega \models p$;*

2 *Markov's inequality holds: $a^2 \cdot \big(\omega \models (\,\big|\, p - (\omega \models p) \cdot \mathbf{1} \,\big| \geq a\,)\big) \leq \mathit{Var}(\omega, p)$.*

Proof. 1 Because:

$$\begin{aligned} \omega \models p \;\geq\; \omega \models p \mathbin{\&} (p \geq a) &\geq \omega \models (a \cdot \mathbf{1}) \mathbin{\&} (p \geq a) \\ &= \omega \models a \cdot (\mathbf{1} \mathbin{\&} (p \geq a)) \\ &= \omega \models a \cdot (p \geq a) \\ &= a \cdot (\omega \models (p \geq a)). \end{aligned}$$

2 Write $q \coloneqq p - (\omega \models p) \cdot \mathbf{1} \colon X \to \mathbb{R}$, so that $q(x) = p(x) - (\omega \models p)$. Then:

$$\begin{aligned}(|q| \geq a)(x) = 1 &\iff |q(x)| \geq a \\ &\implies q(x)^2 \geq a^2 \iff (q^2 \geq a^2)(x) = 1.\end{aligned}$$

This gives an inequality of (sharp) predicates: $(|q| \geq a) \leq (q^2 \geq a^2)$. Hence, by the previous item (Markov's inequality),

$$\begin{aligned}a^2 \cdot \Big(\omega \models \big(\,\big|\, p - (\omega \models p) \cdot \mathbf{1} \,\big|\, \geq a\,\big)\Big) &= a^2 \cdot (\omega \models |q| \geq a) \\ &\leq a^2 \cdot (\omega \models q^2 \geq a^2) \\ &\leq \omega \models q^2 \qquad \text{by item (1)} \\ &= \mathit{Var}(\omega, p).\end{aligned}$$ ❑

We turn to the weak law of large numbers, in binary form. To start, fix the single and parallel coin states:

$$\sigma \coloneqq \tfrac{1}{2}|1\rangle + \tfrac{1}{2}|0\rangle \in \mathcal{D}(2) \qquad \text{and} \qquad \sigma^n \coloneqq \sigma \otimes \cdots \otimes \sigma \in \mathcal{D}(2^n),$$

with a sum predicate on 2^n, namely:

$$2^n \xrightarrow{\mathit{sum}_n} [0, 1] \qquad \text{defined by} \qquad \mathit{sum}_n(x_1, \ldots, x_n) \coloneqq \frac{x_1 + \cdots + x_n}{n}.$$

The predicate sum_n captures the average number of heads (as 1) in n coin flips. Then:

$$\begin{aligned}\sigma \models \mathit{sum}_1 &= \tfrac{1}{2} \cdot 1 + \tfrac{1}{2} \cdot 0 = \tfrac{1}{2} \\ \sigma^2 \models \mathit{sum}_2 &= \frac{\frac{1}{4} \cdot (1 + 1) + \frac{1}{4} \cdot (1 + 0) + \frac{1}{4} \cdot (0 + 1) + \frac{1}{4} \cdot (0 + 0)}{2} = \tfrac{1}{2} \\ &\vdots \\ \sigma^n \models \mathit{sum}_n &= \frac{\sum_{1 \leq k \leq n} \binom{n}{k} \cdot \frac{k}{2^n}}{n} = \tfrac{1}{n} \cdot \mathit{mean}(\mathit{bn}[n](\tfrac{1}{2})) = \tfrac{1}{2}.\end{aligned}$$

For the last equation, see Example 4.1.4 (4). These equations express that in n coin flips the average number of heads is $\frac{1}{2}$. This makes perfect sense.

The weak law of large numbers involves a more subtle statement, namely that for each $\varepsilon > 0$ the validity:

$$\sigma^n \models \big(\,\big|\, \mathit{sum}_n - \tfrac{1}{2} \cdot \mathbf{1} \,\big|\, \geq \varepsilon\,\big). \tag{5.7}$$

goes to zero as n goes to infinity. One may think that this convergence to zero is obvious, but it is not, see Figure 5.2. The above validities (5.7) do go down, but not monotonously.

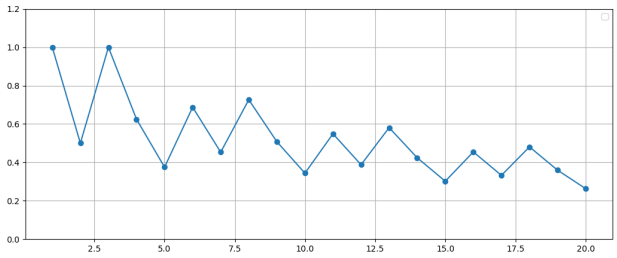


Figure 5.2 Example validities (5.7) with $\varepsilon = \frac{1}{10}$, from $n = 1$ to $n = 20$.

Theorem 5.5.2 (Weak law of large numbers)**.** *Using the fair coin σ and predicate $sum_n \colon 2^n \to [0,1]$ defined above one has, for each $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sigma^n \models \left(\left| sum_n - \tfrac{1}{2} \cdot \mathbf{1} \right| \geq \varepsilon \right) \;=\; 0.$$

Proof. We use Chebyshev's inequality from Lemma 5.5.1 (2). Since $\sigma^n \models sum_n = \frac{1}{2}$, as we have seen above, it gives an inequality in:

$$\sigma^n \models \left(\left| sum_n - \tfrac{1}{2} \cdot \mathbf{1} \right| \geq \varepsilon \right) \leq \frac{Var(\sigma^n, sum_n)}{\varepsilon^2} \overset{(*)}{=} \frac{1}{4n\varepsilon^2}.$$

It suffices to prove the marked equality $\overset{(*)}{=}$, since then it is clear that the limit in the theorem goes to zero, as n goes to infinity.

The product 2^n comes with n projection functions $\pi_i \colon 2^n \to 2$. Using these we can write the predicate $sum_n \colon 2^n \to [0,1]$ as $\frac{1}{n} \cdot \sum_i \pi_i$. Thus:

$$\begin{aligned}
\sigma^n \models sum_n \;\&\; sum_n \;&=\; \frac{1}{n^2} \cdot \sum_{i,j} \sigma^n \models \pi_i \;\&\; \pi_j \\
&= \frac{1}{n^2} \cdot \left(\sum_i \sigma^n \models \pi_i \;\&\; \pi_i + \sum_{i,j,\, i\neq j} \sigma^n \models \pi_i \;\&\; \pi_i \right) \\
&= \frac{1}{n^2} \cdot \left(\frac{n}{2} + \frac{n^2 - n}{4} \right) \\
&= \frac{n+1}{4n}.
\end{aligned}$$

Hence, by Lemma 5.1.3,

$$Var(\sigma^n, sum_n) \;=\; \left(\sigma^n \models sum_n \;\&\; sum_n\right) - \left(\sigma^n \models sum_n\right)^2 \;=\; \frac{n+1}{4n} - \frac{1}{4} \;=\; \frac{1}{4n}.$$

This proves the marked equation $\overset{(*)}{=}$, and thus the theorem. ❑

An obvious next step is to extend this result to arbitrary distributions. So let us fix a distribution ω, with $supp(\omega) = \{x_1, \ldots, x_\ell\} = X$. For a number $n \in \mathbb{N}$ and for $1 \leq i \leq \ell$ we use the accumulation predicate $acc_i \colon X^n \to [0, 1]$, given by:

$$acc_i(y_1, \ldots, y_n) \coloneqq \frac{acc(y_1, \ldots, y_n)(x_i)}{n} = \frac{|\{j \mid y_j = x_i\}|}{n}.$$

Thus, the predicate acc_i yields the average number of occurrences of the element x_i in its input sequence. As one may expect, its value converges to $\omega(x_i)$ in increasing product states $\omega^n = \omega \otimes \cdots \otimes \omega$. This is the content of item (3) below, which the multivariate version of the weak law of large numbers.

Proposition 5.5.3. *Consider the above situation, with a distribution ω and a predicate acc_i.*

1 The validity is given by:

$$\omega^n \models acc_i = \omega(x_i).$$

2 The formula for the variance is:

$$Var(\omega^n, acc_i) = \frac{\omega(x_1) \cdot (1 - \omega(x_i))}{n}.$$

3 For each $\varepsilon > 0$,

$$\lim_{n \to \infty} \omega^n \models \big(\big|\, acc_i - \omega(x_i) \cdot \mathbf{1} \,\big| \geq \varepsilon\big) = 0.$$

Proof. 1 Notice that we can write:

$$acc_i = \tfrac{1}{n} \cdot \Big(X^n \xrightarrow{acc} \mathcal{N}[n](X) \xrightarrow{\pi_{x_i}} \mathbb{R}_{\geq 0}\Big),$$

where π_{x_i} is the point evaluation map from (4.10). Then, using acc as a deterministic channel,

$$\begin{aligned}
\omega^n \models acc_i &= \tfrac{1}{n} \cdot \Big(\omega^n \models acc \mathrel{=\!\!\ll} \pi_{x_i}\Big) \\
&= \tfrac{1}{n} \cdot \Big(acc \gg\!= iid[n](\omega) \models \pi_{x_i}\Big) \\
&= \tfrac{1}{n} \cdot \Big(mn[n](\omega) \models \pi_{x_i}\Big) && \text{by Theorem 3.3.1 (2)} \\
&= \omega(x_i) && \text{by (4.11).}
\end{aligned}$$

2 Via a combination of several earlier results we get:

$$\begin{aligned}
Var\big(\omega^n,\, acc_i\big) &= Var\big(\omega^n,\, \tfrac{1}{n} \cdot (\pi_{x_i} \circ acc)\big) \\
&= \tfrac{1}{n^2} \cdot Var\big(\omega^n,\, \pi_{x_i} \circ acc\big) && \text{by Theorem 5.1.7 (2)} \\
&= \tfrac{1}{n^2} \cdot Var\big(\mathcal{D}(acc)(\omega^n),\, \pi_{x_i}\big) && \text{by Exercise 5.1.5} \\
&= \tfrac{1}{n^2} \cdot Var\big(mn[n](\omega),\, \pi_{x_i}\big) && \text{see Theorem 3.3.1 (2)} \\
&= \frac{\omega(x_1) \cdot (1 - \omega(x_i))}{n} && \text{by Proposition 5.2.1 (1).}
\end{aligned}$$

3 Via Chebyshev's inequality from Lemma 5.5.1 (2), in combination with the previous two items, we get:

$$\begin{aligned}
\lim_{n\to\infty} \omega^n \models \big(\, \big|\, acc_i - \omega(x_i) \cdot \mathbf{1}\, \big| \geq \varepsilon \,\big) &\leq \lim_{n\to\infty} \frac{Var(\omega^n, acc_i)}{\varepsilon^2} \\
&= \lim_{n\to\infty} \frac{\omega(x_1) \cdot (1 - \omega(x_i))}{n \cdot \varepsilon^2} \\
&= 0. \qquad \blacksquare
\end{aligned}$$

In the above description we use the accumulation map $acc\colon X^n \to \mathcal{N}[n](X)$. We shall now write:

$$\frac{acc}{n} \;=\; Flrn \circ acc \;:\; X^n \to \mathcal{D}(X),$$

so that, for instance $\frac{acc}{6}(a, c, a, a, c, b) = \frac{1}{2}|a\rangle + \frac{1}{6}|b\rangle + \frac{1}{3}|c\rangle$.

For a given state $\omega \in \mathcal{D}(X)$ we can look at the total variation distance d, see Subsection 4.5.1, as a predicate $X^n \to [0, 1]$, given by:

$$\vec{x} \longmapsto d\Big(\tfrac{acc(\vec{x})}{n},\, \omega\Big) \;=\; \tfrac{1}{2} \cdot \sum_{y\in X} \Big|\, \tfrac{acc(\vec{x})(y)}{n} - \omega(y)\,\Big|,$$

see Proposition 4.5.1.

This predicate is used in the following multivariate weak large number theorem. Informally it says: the distribution that is obtained by accumulating n samples from ω has a total variation distance to ω that goes to zero as n goes to infinity.

Theorem 5.5.4. *For a state $\omega \in \mathcal{D}(X)$ and a number $\varepsilon > 0$ one has:*

$$\lim_{n\to\infty} \omega^n \models \Big(\, d(\tfrac{acc(-)}{n},\, \omega) \geq \varepsilon \Big) \;=\; 0.$$

Proof. Let $supp(\omega) \subseteq X$ have $\ell \in \mathbb{N}_{>0}$ elements. There is an inequality of sharp predicates on X^n of the form:

$$\sum_{y\in supp(\omega)} \Big(\, \big|\, \tfrac{acc(-)(y)}{n},\, \omega(y)\,\big| \geq \tfrac{2\varepsilon}{\ell} \,\Big) \geq \Big(\, d(\tfrac{acc(-)}{n},\, \omega) \geq \varepsilon \Big). \qquad (*)$$

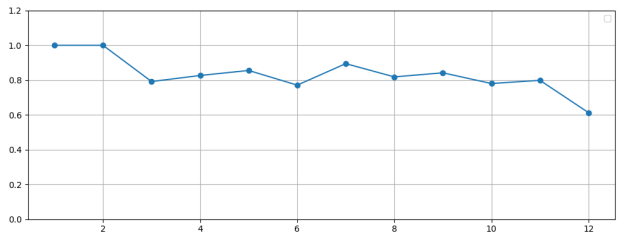


Figure 5.3 Initial segment of the limit from Theorem 5.5.4, with $\omega = \frac{1}{4}|a\rangle + \frac{5}{12}|b\rangle + \frac{1}{3}|c\rangle$ and $\varepsilon = \frac{1}{10}$, from $n = 1$ to $n = 12$.

Indeed, if the left-hand-side in $(*)$ is 0, for $\vec{x} \in X^n$, then for each $y \in \mathit{supp}(\omega)$ we have:

$$\left| \frac{\mathrm{acc}(\vec{x})(y)}{n}, \omega(y) \right| < \frac{\varepsilon}{2\ell}.$$

Hence, by Proposition 5.5.3 (3),

$$d\left(\frac{\mathrm{acc}(\vec{x})}{n}, \omega\right) = \frac{1}{2} \cdot \sum_{y \in \mathit{supp}(\omega)} \left| \frac{\mathrm{acc}(\vec{x})(y)}{n}, \omega(y) \right| < \frac{1}{2} \cdot \sum_{y \in \mathit{supp}(\omega)} \frac{2\varepsilon}{\ell} = \varepsilon.$$

Now we can prove the limit result in the theorem via Proposition 5.5.3 (3):

$$\begin{aligned}
&\lim_{n\to\infty} \omega^n \models \left(d\left(\frac{\mathrm{acc}(-)}{n}, \omega\right) \geq \varepsilon \right) \\
&\leq \lim_{n\to\infty} \omega^n \models \sum_{y \in \mathit{supp}(\omega)} \left(\left| \frac{\mathrm{acc}(-)(y)}{n}, \omega(y) \right| \geq \frac{2\varepsilon}{\ell} \right) \\
&= \sum_{y \in \mathit{supp}(\omega)} \lim_{n\to\infty} \omega^n \models \left(\left| \frac{\mathrm{acc}(-)(y)}{n}, \omega(y) \right| \geq \frac{2\varepsilon}{\ell} \right) \\
&= 0.
\end{aligned}$$

❑

Figure 5.3 gives an impression of the terms in a limit like in Theorem 5.5.4, for $\omega = \frac{1}{4}|a\rangle + \frac{5}{12}|b\rangle + \frac{1}{3}|c\rangle$. The plot covers only a small initial fragment because of the exponential growth of the sample space. For instance, the space of the parallel product ω^{12} has 3^{12} elements, which is more than half a million. This initial segment suggests that the validities go down, but the suggestion is weak; Theorem 5.5.4 provides certainty.

Exercises

5.5.1 For a number $a \in \mathbb{R}$, write $\mathbf{1}_{\geq a} \colon \mathbb{R} \to [0, 1]$ for the sharp predicate with $\mathbf{1}_{\geq a}(r) = 1$ iff $r \geq a$. Consider an observable $p \colon X \to \mathbb{R}$ as a

deterministic channel. Check that the sharp predicate $p \geq a$ from (5.6) can also be described as predicate transformation $p \mathrel{=\!\!\ll} \mathbf{1}_{\geq a}$.

5.5.2 Show how Theorem 5.5.2 is an instance of Theorem 5.5.3 (3).

6

Updating distributions

One of the most interesting and magical aspects of probability distributions (states) is that they can be 'updated'. Informally, this means that in the light of evidence, one can revise a state to a new state, so that the new state better matches the evidence. This updating is also called belief update, conditioning, revision, learning or inference.

Less informally, for a factor (or predicate or event) p and a state ω, both on the same sample space, one can form a new updated (conditioned) state, written as $\omega|_p$. It absorps the evidence p into ω. This updating satisfies various properties, including the famous rule of Thomas Bayes, and also a less well known but also important validity-increase property: $\omega|_p \models p \geq \omega \models p$. It states that after absorbing p the validity of p increases. This makes perfect good sense and is crucial in learning, see Chapter **??**.

Updating is particularly interesting for joint states. A theme that will run through this chapter is that probabilistic updating (conditioning) has 'crossover' influence. This means that if we have a joint state on a product sample space, then updating in one component typically changes the state in the other component (the marginal). This crossover influence is another magical aspect in a probabilistic setting and depends on correlation between the two components. As we shall see, channels play an important role in this phenomenon.

The combination of updating and transformation along a channel adds its own dynamics to the topic. An update $\omega|_p$ involves both a state ω and a factor p. In presence of a channel we can first transform the state ω or the factor p and then update. But we can also transform an update state $\omega|_p$. How are all these related? The power of conditioning becomes apparent when it is combined with transformation, especially for inference in probabilistic reasoning. We shall distinguish two forms of inference: *forward* inference involves conditioning followed by state transformation; it is commonly called causal reasoning. There is also *backward* inference, which is observable transformation

followed by conditioning, and is known as evidential reasoning. We illustrate the usefulness of both these inference techniques in many examples in Sections 6.3 and 6.4, but also for Bayesian networks, in Section 6.5.

Via backward inference we can turn a channel $c\colon X \nrightarrow Y$ into a reversed channel $c^\dagger\colon Y \nrightarrow X$, which is commonly called the *dagger* of c. The construction also involves a state on X, but it is left out at this stage for convenience. This reversed channel plays an important role in Theorem 6.6.3, which is one of the more elegant results in Bayesian updating: it relates forward (resp. backward) inference along a channel c to backward (resp. forward) inference along the reversed channel $c^\dagger$. In this chapter we first introduce dagger channels for probabilistic reasoning. A more thorough analysis appears in Chapter 7, using the graphical language of string diagrams.

A special topic is backward inference, which is a form of updating along a channel. It arises in a situation with a channel $c\colon X \nrightarrow Y$ with a state $\sigma \in \mathcal{D}(X)$ on its domain. If we have evidence on the codomain Y, we would like to update σ after transporting the evidence along the channel c. It turns out that there are two fundamentally different mechanisms for doing so.

1 In the first case the evidence is given in the form of a *predicate* (or factor) on Y. One can then apply backward inference. This is called Pearl's rule, after [135] (and Bayes).
2 In the second case the evidence is given by a *distribution* on Y. One can then do state transformation with the inverted channel (dagger) $c^\dagger$. This is called Jeffrey's rule, after [90].

In general, these two update mechanisms produce different outcomes. This is uncomfortable, especially because it is not clear under which circumstances one should use which update rule. The mathematical characterisation of these update rules gives some guidance. Briefly, Pearl's update rule gives an increase of validity and Jeffrey's update rule yields a decrease of divergence. This will be made mathematically precise in Section 6.7 below. Informally, this involves the difference between learning as reinforcing what goes well (go higher), or learning as steering away from what goes wrong (go less low).

6.1 Update basics

We shall use updating and conditioning synonymously. These terms refer to the incorporation of evidence into a distribution, where the evidence is given by a predicate (or more generally by a factor). This section describes the definition and basic results, including Bayes' theorem and validity-increase. The

relevance of conditioning in probabilistic reasoning will be demonstrated in many examples later on in this chapter.

Definition 6.1.1. Let $\omega \in \mathcal{D}(X)$ be a distribution on a sample space X and let $p \in \mathit{Fact}(X) = (\mathbb{R}_{\geq 0})^X$ be a factor, on the same space X.

1 If the validity $\omega \models p$ is non-zero, we define a new distribution $\omega|_p \in \mathcal{D}(X)$ as normalised pointwise product of ω and p:

$$\omega|_p(x) \coloneqq \frac{\omega(x) \cdot p(x)}{\omega \models p} \qquad \textit{i.e.} \qquad \omega|_p = \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \models p} \big|x\big\rangle.$$

This distribution $\omega|_p$ may be pronounced as: ω, given p.

2 The *conditional expectation* or *conditional validity* of an observable q on X, given p and ω, is the validity:

$$\omega|_p \models q.$$

3 For a channel $c \colon X \nrightarrow Y$ and a factor q on Y we define the updated channel $c|_q \colon X \nrightarrow Y$ via pointwise updating:

$$c|_q(x) \coloneqq c(x)|_q.$$

In writing $c|_q$ we assume that validity $c(x) \models q = (c \mathrel{=\!\!\ll} q)(x)$ is non-zero, for each $x \in X$.

Often, the state ω before updating is called the *prior* or the *a priori* state, whereas the updated state $\omega|_p$ is called the *posterior* or the *a posteriori* state. The posterior incorporates the evidence given by the factor p. One may thus expect that in the updated state $\omega|_p$ the factor p is more true than in ω. This is indeed the case, as will be shown in Theorem 6.1.5 below.

The conditioning $c|_q$ of a channel in item (3) is in fact a generalisation of the conditioning of a state $\omega|_p$ in item (1), since the state $\omega \in \mathcal{D}(X)$ can be seen as a channel $\omega \colon 1 \nrightarrow X$ with a one-element set $1 = \{0\}$ as domain. We shall see the usefulness of conditioning of channels in Subsection 6.3.1.

The standard conditional probability notation is: $P(E \mid D)$ for events $E, D \subseteq X$, where the distribution involved is left implicitly. If ω is this implicit distribution, then $P(E \mid D)$ corresponds to the conditional expectation expressed by the validity $\omega|_{\mathbf{1}_D} \models \mathbf{1}_E$ of the sharp predicate $\mathbf{1}_E$ in the state ω updated with the

sharp predicate $\mathbf{1}_D$. Indeed,

$$\begin{aligned}\omega|_{\mathbf{1}_D} \models \mathbf{1}_E &= \sum_x \omega|_{\mathbf{1}_D}(x) \cdot \mathbf{1}_E(x) \\ &= \sum_{x \in E} \frac{\omega(x) \cdot \mathbf{1}_D(x)}{\omega \models \mathbf{1}_D} \\ &= \frac{\sum_{x \in E \cap D} \omega(x)}{\sum_{x \in D} \omega(x)} = \frac{P(E \cap D)}{P(D)} = P(E \mid D).\end{aligned}$$

The formulation $\omega|_p$ of conditioning that is used above is not restricted to sharp predicates, but works much more generally for fuzzy predicates/factors p. This is sometimes called updating with *soft* or *uncertain* evidence [19, 33, 77]. It is what we use right from the start.

Example 6.1.2. 1 Let's take the numbers of a dice as sample space: $\mathit{pips} = \{1, 2, 3, 4, 5, 6\}$, with a fair/uniform dice distribution $\mathit{dice} = \mathit{unif}_{\mathit{pips}} = \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$. We consider the predicate $\mathit{evenish} \in \mathit{Pred}(\mathit{pips}) = [0, 1]^{\mathit{pips}}$ expressing that we are fairly certain of the outcome being even:

$$\begin{array}{lll} \mathit{evenish}(1) = \frac{1}{5} & \mathit{evenish}(3) = \frac{1}{10} & \mathit{evenish}(5) = \frac{1}{10} \\ \mathit{evenish}(2) = \frac{9}{10} & \mathit{evenish}(4) = \frac{9}{10} & \mathit{evenish}(6) = \frac{4}{5} \end{array}$$

We first compute the validity of *evenish* for our fair dice:

$$\begin{aligned}\mathit{dice} \models \mathit{evenish} &= \textstyle\sum_x \mathit{dice}(x) \cdot \mathit{evenish}(x) \\ &= \tfrac{1}{6} \cdot \tfrac{1}{5} + \tfrac{1}{6} \cdot \tfrac{9}{10} + \tfrac{1}{6} \cdot \tfrac{1}{10} + \tfrac{1}{6} \cdot \tfrac{9}{10} + \tfrac{1}{6} \cdot \tfrac{1}{10} + \tfrac{1}{6} \cdot \tfrac{4}{5} \\ &= \tfrac{2+9+1+9+1+8}{60} = \tfrac{1}{2}.\end{aligned}$$

If we take *evenish* as evidence, we can update our *dice* state and get:

$$\begin{aligned}\mathit{dice}\big|_{\mathit{evenish}} &= \sum_x \frac{\mathit{dice}(x) \cdot \mathit{evenish}(x)}{\mathit{dice} \models \mathit{evenish}} \big|x\rangle \\ &= \frac{1/6 \cdot 1/5}{1/2}\big|1\big\rangle + \frac{1/6 \cdot 9/10}{1/2}\big|2\big\rangle + \frac{1/6 \cdot 1/10}{1/2}\big|3\big\rangle \\ &\qquad + \frac{1/6 \cdot 9/10}{1/2}\big|4\big\rangle + \frac{1/6 \cdot 1/10}{1/2}\big|5\big\rangle + \frac{1/6 \cdot 4/5}{1/2}\big|6\big\rangle \\ &= \tfrac{1}{15}\big|1\big\rangle + \tfrac{3}{10}\big|2\big\rangle + \tfrac{1}{30}\big|3\big\rangle + \tfrac{3}{10}\big|4\big\rangle + \tfrac{1}{30}\big|5\big\rangle + \tfrac{4}{15}\big|6\big\rangle.\end{aligned}$$

As expected, the probabilities for the even pips is now, in the posterior, higher than for the odd ones.

2 The following alarm example is due to Pearl [133]. It involves an 'alarm' set $A = \{a, a^\perp\}$ and a 'burglary' set $B = \{b, b^\perp\}$, with the following a priori joint distribution $\omega \in \mathcal{D}(A \times B)$.

$$0.000095|a, b\rangle + 0.009999|a, b^\perp\rangle + 0.000005|a^\perp, b\rangle + 0.989901|a^\perp, b^\perp\rangle.$$

The *a priori* burglary distribution is the second marginal:

$$\omega[0,1] \;=\; 0.0001|b\rangle + 0.9999|b^{\perp}\rangle.$$

Someone reports that the alarm went off, but with only 80% certainty because of deafness. This can be described as a predicate $p\colon A \to [0,1]$ with $p(a) = 0.8$ and $p(a^{\perp}) = 0.2$. We see that there is a mismatch between the p on A and ω on $A \times B$. This can be solved via weakening p to $p \otimes \mathbf{1}$, so that it becomes a predicate on $A \times B$. Then we can take the update:

$$\omega|_{p\otimes\mathbf{1}} \;\in\; \mathcal{D}(A \times B).$$

In order to do so we first compute the validity:

$$\omega \models p \otimes \mathbf{1} \;=\; \sum\nolimits_{x,y} \omega(x,y) \cdot p(x) \;=\; 0.206.$$

We can then compute the updated joint state as:

$$\begin{aligned}\omega|_{p\otimes\mathbf{1}} \;&=\; \sum\nolimits_{x,y} \frac{\omega(x,y) \cdot p(x)}{\omega \models p} \,\big|x,y\big\rangle \\ &= 0.0003688|a,b\rangle + 0.03882|a,b^{\perp}\rangle \\ &\qquad + 0.000004853|a^{\perp},b\rangle + 0.9608|a^{\perp},b^{\perp}\rangle.\end{aligned}$$

The resulting posterior burglary distribution — with the alarm evidence taken into account — is obtained by taking the second marginal of the updated distribution:

$$(\omega|_{p\otimes\mathbf{1}})[0,1] \;=\; 0.0004|b\rangle + 0.9996|b^{\perp}\rangle.$$

We see that the burglary probability is four times higher in the posterior than in the prior. What happens is noteworthy: evidence about one component A changes the probabilities in another component B. This 'crossover influence' (in the terminology of [88]) or 'crossover inference' happens precisely because the joint distribution ω is entwined, so that the different parts can influence each other. We shall return to this theme repeatedly, see for instance in Corollary 6.3.13 below.

One of the main results about conditioning is Bayes' theorem. We present it here for factors, and not just for sharp predicates (events), as is common.

Theorem 6.1.3. *Let ω be distribution on a sample space X, and let p, q be factors on X.*

1 The product rule holds:

$$\omega|_p \models q \;=\; \frac{\omega \models p \,\&\, q}{\omega \models p}.$$

2 Bayes' rule holds:

$$\omega|_p \models q \;=\; \frac{(\omega|_q \models p) \cdot (\omega \models q)}{\omega \models p}.$$

This result carefully distinguishes a product rule, in item 1, and Bayes' rule, in item (2). This distinction is not always made, and the product rule is sometimes also called Bayes' rule.

Proof. 1 We straightforwardly compute:

$$\begin{aligned}\omega|_p \models q \;=\; \sum\nolimits_x \omega|_p(x) \cdot q(x) \;&=\; \sum\nolimits_x \frac{\omega(x) \cdot p(x)}{\omega \models p} \cdot q(x) \\ &=\; \frac{\sum_x \omega(x) \cdot p(x) \cdot q(x)}{\omega \models p} \\ &=\; \frac{\sum_x \omega(x) \cdot (p \,\&\, q)(x)}{\omega \models p} \;=\; \frac{\omega \models p \,\&\, q}{\omega \models p}.\end{aligned}$$

2 This follows directly by using the previous item twice, in combination with the commutativity of &, in:

$$\omega|_p \models q \overset{(1)}{=} \frac{\omega \models p \,\&\, q}{\omega \models p} \;=\; \frac{\omega \models q \,\&\, p}{\omega \models p} \overset{(1)}{=} \frac{(\omega|_q \models p) \cdot (\omega \models q)}{\omega \models p}. \qquad \text{❑}$$

Example 6.1.4. We instantiate Proposition 6.1.3 with sharp predicates $\mathbf{1}_E, \mathbf{1}_D$ for subsets/events $E, D \subseteq X$. Then the familiar formulations of the product/Bayes rule appear.

1 The product rule specialises to the definition of conditional probability:

$$P(E \mid D) \;=\; \omega|_{\mathbf{1}_D} \models \mathbf{1}_E \;=\; \frac{\omega \models \mathbf{1}_D \,\&\, \mathbf{1}_E}{\omega \models \mathbf{1}_D} \;=\; \frac{\omega \models \mathbf{1}_{D \cap E}}{P(D)} \;=\; \frac{P(D \cap E)}{P(D)}.$$

2 Bayes rule, in the general formulation of Proposition 6.1.3 (2) specialises to the well known inversion property of conditional probabilities:

$$P(E \mid D) \;=\; \omega|_{\mathbf{1}_D} \models \mathbf{1}_E \;=\; \frac{(\omega|_{\mathbf{1}_E} \models \mathbf{1}_D) \cdot (\omega \models \mathbf{1}_D)}{\omega \models \mathbf{1}_E} \;=\; \frac{P(D \mid E) \cdot P(E)}{P(D)}.$$

We have explained updating $\omega|_p$ as incorporating the evidence p into the distribution ω. Thus, one expects p to be 'more true' in $\omega|_p$ than in ω. The next result shows that this is indeed the case. It plays an important role in learning see Chapter **??**.

Theorem 6.1.5 (Validity-increase)**.** *For a distribution ω and a factor p on the same set, if the validity $\omega \models p$ is non-zero, one has:*

$$\omega|_p \models p \;\geq\; \omega \models p.$$

Proof. We recall the 'byproduct' $\omega \models p \mathbin{\&} p \geq (\omega \models p)^2$ from Lemma 5.1.3. Then, by the product rule from Theorem 6.1.3 (1),

$$\omega|_p \models p = \frac{\omega \models p \mathbin{\&} p}{\omega \models p} \geq \frac{(\omega \models p)^2}{\omega \models p} = \omega \models p. \qquad \blacksquare$$

We add a few more basic facts about conditioning.

Lemma 6.1.6. *Let ω be distribution on X, with factors $p, q \in \mathrm{Fact}(X)$.*

1 Conditioning with truth has no effect:

$$\omega|_{\mathbf{1}} = \omega.$$

2 Conditioning with a point predicate gives a point state: for $a \in X$,

$$\omega|_{\mathbf{1}_a} = 1|a\rangle, \qquad \text{assuming } \omega(a) \neq 0.$$

3 Iterated conditionings commute:

$$(\omega|_p)|_q = \omega|_{p \& q} = (\omega|_q)|_p.$$

4 Conditioning is stable under multiplication of the factor with a scalar $s > 0$:

$$\omega|_{s \cdot p} = \omega|_p.$$

5 Conditioning can be done component-wise, for product states and factors:

$$(\sigma \otimes \tau)\big|_{(p \otimes q)} = (\sigma|_q) \otimes (\tau|_p).$$

6 Marginalisation of a conditioning with a (similarly) weakened predicate is conditioning of the marginalised state:

$$\omega|_{\mathbf{1} \otimes q}[0, 1] = \omega[0, 1]\big|_q.$$

7 For a function $f \colon X \to Y$, used as deterministic channel,

$$(f \gg\!= \omega)\big|_q = f \gg\!= (\omega\big|_{q \circ f}).$$

When we ignore undefinedness issues, we see that items 1 and 3 show that conditioning is an action on distributions, namely of the monoid of factors with conjunction $(\mathbf{1}, \&)$, see Definition 1.2.4.

Proof. 1 Trivial since $\omega \models \mathbf{1} = 1$.

2 Assuming $\omega(a) \neq 0$ we get for each $x \in X$,

$$\omega|_{\mathbf{1}_a}(x) = \frac{\omega(x) \cdot \mathbf{1}_a(x)}{\omega \models \mathbf{1}_a} = \frac{\omega(a) \cdot 1|a\rangle(x)}{\omega(a)} = 1|a\rangle(x).$$

3 It suffices to prove:

$$(\omega|_p)|_q(x) \;=\; \frac{\omega|_p(x)\cdot q(x)}{\omega|_p \models q} \;=\; \frac{{}^{\omega(x)\cdot p(x)}/_{\omega\models p}\cdot q(x)}{{}^{\omega\models p\&q}/_{\omega\models p}} \quad \text{by Proposition 6.1.3 (1)}$$
$$= \frac{\omega(x)\cdot(p\mathbin{\&} q)(x)}{\omega\models p\mathbin{\&} q} \;=\; \omega|_{p\&q}(x).$$

4 First we have:

$$\omega\models s\cdot p \;=\; \sum\nolimits_x \omega(x)\cdot(s\cdot p)(x) \;=\; \sum\nolimits_x \omega(x)\cdot s\cdot p(x)$$
$$= s\cdot\sum\nolimits_x \omega(x)\cdot p(x) \;=\; s\cdot(\omega\models p).$$

Next:

$$\omega|_{s\cdot p}(x) \;=\; \frac{\omega(x)\cdot(s\cdot p)(x)}{\omega\models s\cdot p} \;=\; \frac{\omega(x)\cdot s\cdot p(x)}{s\cdot(\omega\models p)} \;=\; \frac{\omega(x)\cdot p(x)}{\omega\models p} \;=\; \omega|_p(x).$$

5 For states $\sigma\in\mathcal{D}(X)$, $\tau\in\mathcal{D}(Y)$ and factors p on X and q on Y one has:

$$\Big((\sigma\otimes\tau)\big|_{(p\otimes q)}\Big)(x,y) \;=\; \frac{(\sigma\otimes\tau)(x,y)\cdot(p\otimes q)(x,y)}{(\sigma\otimes\tau)\models(p\otimes q)}$$
$$= \frac{\sigma(x)\cdot\tau(y)\cdot p(x)\cdot q(y)}{(\sigma\models p)\cdot(\tau\models q)} \qquad \text{by Lemma 4.2.8}$$
$$= \frac{\sigma(x)\cdot p(x)}{\sigma\models p}\cdot\frac{\tau(y)\cdot q(y)}{\tau\models q}$$
$$= (\sigma|_p)(x)\cdot(\tau|_q)(y)$$
$$= \Big((\sigma|_p)\otimes(\tau|_q)\Big)(x,y).$$

6 Let $\omega\in\mathcal{D}(X\times Y)$ and q be a factor on Y; then:

$$(\omega|_{\mathbf{1}\otimes q}[0,1])(y) \;=\; \sum\nolimits_x \omega|_{\mathbf{1}\otimes q}(x,y)$$
$$= \sum\nolimits_x \frac{\omega(x,y)\cdot(\mathbf{1}\otimes q)(x,y)}{\omega\models\mathbf{1}\otimes q}$$
$$\overset{(4.8)}{=} \frac{(\omega[0,1])(y)\cdot q(y)}{\omega[0,1]\models q} \;=\; (\omega[0,1]\big|_q)(y).$$

7 For $f\colon X\to Y$, $\omega\in\mathcal{D}(X)$ and $q\in \mathit{Fact}(Y)$,

$$((f\gg\!= \omega)|_q)(y) \;=\; \mathcal{D}(f)(\omega)|_q(y) \;=\; \frac{\mathcal{D}(f)(\omega)(y)\cdot q(y)}{\mathcal{D}(f)(\omega)\models q}$$
$$= \sum\nolimits_{x\in f^{-1}(y)} \frac{\omega(x)\cdot q(y)}{\sum_{x\in f^{-1}(y)}\omega(x)\cdot p(y)}$$
$$= \sum\nolimits_{x\in f^{-1}(y)} \frac{\omega(x)\cdot q(f(x))}{\sum_x \omega(x)\cdot q(f(x))}$$
$$= \sum\nolimits_{x\in f^{-1}(y)} \omega|_{q\circ f}(x)$$
$$= \mathcal{D}(f)(\omega|_{q\circ f})(y) \;=\; (f\gg\!= (\omega|_{q\circ f}))(y). \quad \text{❑}$$

In the beginning of this section we have defined updating $\omega|_p$ for a state $\omega \in \mathcal{D}(X)$ and a *factor* $p\colon X \to \mathbb{R}_{\geq 0}$. Now let's assume that this factor p is bounded: there is a bound $B \in \mathbb{R}_{>0}$ such that $p(x) \leq B$ for all $x \in X$. The rescaled factor $\frac{1}{B} \cdot p$ is then a predicate. Proposition 6.1.6 (4) shows that updating with the factor p is the same as updating with the predicate $\frac{1}{B} \cdot p$. Hence we do not loose much if we restrict updating to predicates. Nevertheless it is most convenient to define updating for factors so that we do not have to bother about any rescaling.

Exercises

6.1.1 Let $\omega \in \mathcal{D}(X)$ with $a \in \mathit{supp}(\omega)$. Prove that:

$$\omega \models \mathbf{1}_a = \omega(x) \qquad \text{and} \qquad \omega|_{\mathbf{1}_a} = 1|a\rangle.$$

6.1.2 Consider the following girls/boys riddle: given that a family with two children has a boy, what is the probability that the other child is a girl? Take as space $\{G, B\}$. On it we use the uniform distribution $\mathit{unif} = \frac{1}{2}|G\rangle + \frac{1}{2}|B\rangle$ since there is no prior knowledge.

1 Take as 'at least one girl' and 'at least one boy' predicates on $\{G, B\} \times \{G, B\}$:

$$g \coloneqq (\mathbf{1}_B \otimes \mathbf{1}_B)^\perp = (\mathbf{1}_G \otimes \mathbf{1}) \otimes (\mathbf{1}_B \otimes \mathbf{1}_G)$$
$$b \coloneqq (\mathbf{1}_G \otimes \mathbf{1}_G)^\perp = (\mathbf{1}_B \otimes \mathbf{1}) \otimes (\mathbf{1}_G \otimes \mathbf{1}_B).$$

Compute $\mathit{unif} \otimes \mathit{unif} \models g$ and $\mathit{unif} \otimes \mathit{unif} \models b$.

2 Check that $\mathit{unif} \otimes \mathit{unif}\big|_b \models g = \frac{2}{3}$.

(For a description and solution of this problem in a special library for probabilistic programming of the functional programming language Haskell, see [45].)

6.1.3 In the setting of Example 6.1.2 (1) define a new predicate $\mathit{oddish} = \mathit{evenish}^\perp = \mathbf{1} - \mathit{evenish}$.

1 Compute $\mathit{dice}|_{\mathit{oddish}}$

2 Prove the equation below, involving a convex sum of states on the left-hand side.

$$(\mathit{dice} \models \mathit{evenish}) \cdot \omega|_{\mathit{evenish}} + (\mathit{dice} \models \mathit{oddish}) \cdot \omega|_{\mathit{oddish}} = \mathit{dice}.$$

6.1.4 Let $p_1, \ldots, p_n \in \mathit{Pred}(X)$ be a test, *i.e.* an n-tuple of predicates on X with $p_1 \otimes \cdots \otimes p_n = \mathbf{1}$. Let $\omega \in \mathcal{D}(X)$ and $q \in \mathit{Fact}(X)$.

1 Check that $1 = \sum_i (\omega \models p_i)$.

2 Prove what is called the *law of total probability*:

$$\omega \;=\; \sum_i (\omega \models p_i) \cdot \omega|_{p_i}. \tag{6.1}$$

What happens to the expression on the right-hand side if one of the p_i has validity zero? Check that this equation generalises Exercise 6.1.3.

(The expression on the right-hand side in (6.1) is used to turn a test into a 'denotation' function $\mathcal{D}(X) \to \mathcal{D}(\mathcal{D}(X))$ in [121, 122], namely as $\omega \mapsto \sum_i (\omega \models p_i) \big|\omega|_{p_i}\big\rangle$. This proces is described more abstractly in terms of 'hypernormalisation' in [72].)

3 Show that:

$$\omega \models q \;=\; \sum_i \omega \models q \mathbin{\&} p_i.$$

4 Prove now:

$$\omega|_q \models p_i \;=\; \frac{\omega \models q \mathbin{\&} p_i}{\sum_j \omega \models q \mathbin{\&} p_j}.$$

6.1.5 Show that conditioning a convex sum of states yields another convex sum of conditioned states: for $\sigma, \tau \in \mathcal{D}(X)$, $p \in \mathit{Fact}(X)$ and $r, s \in [0, 1]$ with $r + s = 1$,

$$\begin{aligned}
&(r \cdot \sigma + s \cdot \tau)\big|_p \\
&\quad= \frac{r \cdot (\sigma \models p)}{r \cdot (\sigma \models p) + s \cdot (\tau \models p)} \cdot \sigma|_p \;+\; \frac{s \cdot (\tau \models p)}{r \cdot (\sigma \models p) + s \cdot (\tau \models p)} \cdot \tau|_p \\
&\quad= \frac{r \cdot (\sigma \models p)}{(r \cdot \sigma + s \cdot \tau) \models p} \cdot \sigma|_p \;+\; \frac{s \cdot (\tau \models p)}{(r \cdot \sigma + s \cdot \tau) \models p} \cdot \tau|_p.
\end{aligned}$$

6.1.6 Consider $\omega \in \mathcal{D}(X)$ and $p \in \mathit{Fact}(X)$ where p is non-zero, at least on the support of X. Check that updating ω with p can be undone via updating with $\frac{1}{p}$.

6.1.7 Let $c \colon X \nrightarrow Y$ be a channel, with state $\omega \in \mathcal{D}(X \times Z)$.

1 Prove that for a factor $p \in \mathit{Fact}(Z)$,

$$((c \otimes id) \gg\!= \omega)\big|_{\mathbf{1} \otimes p} \;=\; (c \otimes id) \gg\!= (\omega\big|_{\mathbf{1} \otimes p}).$$

2 Show also that for $q \in \mathit{Fact}(Y \times Z)$,

$$\Big(((c \otimes id) \gg\!= \omega)\big|_q\Big)[0, 1] \;=\; \Big(\omega\big|_{(c \otimes id) \mathrel{=\!\!\ll} q}\Big)[0, 1].$$

6.1.8 This exercise will demonstrate that conditioning may both create and remove entwinedness.

1 Write $yes = \mathbf{1}_1 \colon 2 \to [0,1]$, where $2 = \{0,1\}$, and $no = yes^{\perp} = \mathbf{1}_0$. Prove that the following conditioning of a non-entwined state,

$$\tau \coloneqq (flip \otimes flip)\big|_{(yes \otimes yes) \oplus (no \otimes no)}$$

is entwined.

2 Consider the state $\omega \in \mathcal{D}(2 \times 2 \times 2)$ given by:

$$\begin{aligned}\omega = \tfrac{1}{18}|111\rangle + \tfrac{1}{9}|110\rangle + \tfrac{2}{9}|101\rangle + \tfrac{1}{9}|100\rangle \\ + \tfrac{1}{9}|011\rangle + \tfrac{2}{9}|010\rangle + \tfrac{1}{9}|001\rangle + \tfrac{1}{18}|000\rangle\end{aligned}$$

Prove that ω's first and third component are entwined:

$$\omega[1,0,1] \neq \omega[1,0,0] \otimes \omega[0,0,1].$$

3 Now let ρ be the following conditioning of ω:

$$\rho \coloneqq \omega|_{\mathbf{1} \otimes yes \otimes \mathbf{1}}.$$

Prove that ρ's first and third component are non-entwined.

The phenomenon that entwined states become non-entwined via conditioning is called *screening-off*, whereas the opposite, non-entwined states becoming entwined via conditioning, is called *explaining away*.

6.1.9 Show that for $\omega \in \mathcal{D}(X)$ and $p_1, p_2 \in Fact(X)$ one has:

$$(\Delta \gg\!= \omega)\big|_{p_1 \otimes p_2} = \Delta \gg\!= (\omega|_{p_1 \& p_2}).$$

Note that this is a consequence of Lemma 6.1.6 (7).

6.1.10 We have mentioned (right after Definition 6.1.1) that updating of the identity channel has no effect. Prove more generally that for an ordinary function $f \colon X \to Y$, updating the associated deterministic channel $\langle f \rangle \colon X \to\!\!\!\!\bullet\; Y$ has no effect:

$$\langle f \rangle|_q = \langle f \rangle.$$

6.1.11 Let $c \colon Z \to\!\!\!\!\bullet\; X$ and $d \colon Z \to\!\!\!\!\bullet\; Y$ be two channels with a common domain Z, and with factors $p \in Fact(X)$ and $q \in Fact(Y)$ on their codomains. Prove that the update of a tuple channel is the tuple of the updates:

$$\langle c, d \rangle|_{p \otimes q} = \langle c|_p, d|_q \rangle.$$

Prove also that for $e \colon U \to\!\!\!\!\bullet\; X$ and $f \colon V \to\!\!\!\!\bullet\; Y$,

$$(e \otimes f)|_{p \otimes q} = (e|_p) \otimes (f|_q).$$

6.1.12 Consider a state ω and predicate p on the same set. For $r \in [0,1]$ we use the (*ad hoc*) abbreviation ω_r for the convex combination of updated states:

$$\omega_r \coloneqq r \cdot \omega|_p + (1-r) \cdot \omega|_{p^\perp}.$$

1 Prove that:

$$r \leq (\omega \models p) \Longrightarrow r \leq (\omega_r \models p) \leq (\omega \models p)$$
$$r \geq (\omega \models p) \Longrightarrow r \geq (\omega_r \models p) \geq (\omega \models p).$$

2 Show that, still for all $r \in [0,1]$,

$$r \cdot \frac{\omega \models p}{\omega_r \models p} + (1-r) \cdot \frac{\omega \models p^\perp}{\omega_r \models p^\perp} \leq 1.$$

Hint: Write $1 = r + (1-r)$ on the right-hand side of the inequality $\leq$ and move the two summands to the left.
This inequality occurs in n-ary form in Proposition 6.7.10 below. It is due to [50].

6.2 Updating draw distributions

In this section we apply updating to distributions that are obtained by drawing from an urn, as described in Chapter 3. We first show that conditioning commutes with multinomial channels. The proof is not particularly difficult, but the result is noteworthy because it shows how well basic probabilistic notions are integrated.

Recall that $\overline{p}^\bullet \colon \mathcal{N}(X) \to \mathbb{R}$ is the multiplicative extension of an observable $p \colon X \to \mathbb{R}$, see Definition 4.4.2. In Proposition 4.4.3 (2) we already saw that the validity $mn[K](\omega) \models \overline{p}^\bullet$ in a multinomial distribution equals the K-power $(\omega \models p)^K$ of the validity in the original distribution. We now extend this result to updating.

Theorem 6.2.1. *Let $\omega \in \mathcal{D}(X)$ be given. For a factor (or predicate) $p \colon X \to \mathbb{R}_{\geq 0}$ and a number $K \in \mathbb{N}$,*

1 $mn[K](\omega)\big|_{\overline{p}^\bullet} = mn[K](\omega|_p)$;
2 Similarly, for the parallel multinomial law pml *from Section 3.6,*

$$pml(\textstyle\sum_i n_i|\omega_i\rangle)\big|_{\overline{p}^\bullet} = pml(\textstyle\sum_i n_i|\,\omega_i|_p\rangle).$$

An example that combines multinomials and updating occurs in [145, §6.4], but without a general formulation.

Proof. 1 Via the validity formula of Proposition 4.4.3 (2):

$$\begin{aligned}
mn[K](\omega)\big|_{\overline{p}^\bullet} &= \sum_{\varphi \in \mathcal{N}[K](X)} \frac{mn[K](\omega)(\varphi) \cdot \overline{p}^\bullet(\varphi)}{mn[K](\omega) \models \overline{p}^\bullet} \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \binom{K}{\varphi} \cdot \frac{\prod_x (\omega(x) \cdot p(x))^{\varphi(x)}}{(\omega \models p)^K} \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \binom{K}{\varphi} \cdot \prod\nolimits_x \left(\frac{\omega(x)}{\omega \models p}\right)^{\varphi(x)} \qquad \text{since } \|\varphi\| = K \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \binom{K}{\varphi} \cdot \prod\nolimits_x \omega|_p(x)^{\varphi(x)} \\
&= mn[K](\omega|_p).
\end{aligned}$$

2 Similarly, via Proposition 4.4.5:

$$\begin{aligned}
&pml(\textstyle\sum_i n_i|\omega_i\rangle)\big|_{\overline{p}^\bullet} \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \frac{pml(\sum_i n_i|\omega_i\rangle)(\varphi) \cdot \overline{p}^\bullet(\varphi)}{pml(\sum_i n_i|\omega_i\rangle) \models \overline{p}^\bullet} \big|\varphi\rangle \quad \text{where } K = \textstyle\sum_i n_i \\
&\overset{(3.25)}{=} \sum_{i, \varphi_i \in \mathcal{N}[n_i](X)} \frac{\prod_i mn[n_i](\omega_i)(\varphi_i) \cdot \overline{p}^\bullet(\varphi_i)}{\prod_i (\omega_i \models p)^{n_i}} \big| \textstyle\sum_i \varphi_i \big\rangle \\
&= \sum_{i, \varphi_i \in \mathcal{N}[n_i](X)} \prod\nolimits_i \binom{n_i}{\varphi_i} \cdot \frac{\prod_x (\omega(x) \cdot p(x))^{\varphi_i(x)}}{(\omega_i \models p)^{n_i}} \big| \textstyle\sum_i \varphi_i \big\rangle \\
&= \sum_{i, \varphi_i \in \mathcal{N}[n_i](X)} \prod\nolimits_i \binom{n_i}{\varphi_i} \cdot \prod\nolimits_x \left(\frac{\omega(x) \cdot p(x)}{\omega_i \models p}\right)^{\varphi_i(x)} \big| \textstyle\sum_i \varphi_i \big\rangle \\
&= \sum_{i, \varphi_i \in \mathcal{N}[n_i](X)} \prod\nolimits_i \binom{n_i}{\varphi_i} \cdot \prod\nolimits_x \omega_i|_p(x)^{\varphi_i(x)} \big| \textstyle\sum_i \varphi_i \big\rangle \\
&= \sum_{i, \varphi_i \in \mathcal{N}[n_i](X)} \prod\nolimits_i mn[n_i](\omega_i|_p)(\varphi_i) \big| \textstyle\sum_i \varphi_i \big\rangle \\
&= pml(\textstyle\sum_i n_i|\, \omega_i|_p \rangle).
\end{aligned}$$

❑

The action $\varphi \bullet p$ of a factor p on a multiset φ from Exercise 4.2.13 can be used to get an analogue of the first item in the previous proposition for hypergeometric and Pólya distributions. However, this result is more restricted and works only for a *sharp* predicate p. The sharp predicate restricts both the urn and the draws from it to balls of certain colours only.

Theorem 6.2.2. *Let $\psi \in \mathcal{N}(X)$ be an urn and $E \subseteq X$ an event/subset, forming a sharp indicator predicate $\mathbf{1}_E \colon X \to [0,1]$.*

1 If $\left\| \psi \bullet \mathbf{1}_E \right\| \geq K$,

$$hg[K](\psi)\big|_{\overline{\mathbf{1}_E}^{\bullet}} \;=\; hg[K]\big(\psi \bullet \mathbf{1}_E\big).$$

2 Similarly, if $\psi \bullet \mathbf{1}_E$ *is non-empty,*

$$pl[K](\psi)\big|_{\overline{\mathbf{1}_E}^{\bullet}} \;=\; pl[K](\psi \bullet \mathbf{1}_E).$$

Proof. 1 We first notice that for $\varphi \in \mathcal{N}(X)$,

$$\begin{aligned} \overline{\mathbf{1}_E}^{\bullet}(\varphi) \;=\; 1 &\iff \prod_{x \in X} \big(\mathbf{1}_E(x)\big)^{\varphi(x)} \;=\; 1 \\ &\iff \forall x \in X.\, x \notin E \Rightarrow \varphi(x) = 0 \\ &\iff supp(\varphi) \subseteq E. \end{aligned}$$

Next, let's write $L = \|\psi\|$ for the size of the urn and $L_E = \|\psi \bullet \mathbf{1}_E\|$ for the size of the restricted urn — where it is assumed that $L_E \geq K$.

$$\begin{aligned} hg[K](\psi) \models \overline{\mathbf{1}_E}^{\bullet} \;&=\; \sum_{\varphi \leq_K \psi} hg[K](\psi)(\varphi) \cdot \overline{\mathbf{1}_E}^{\bullet}(\varphi) \\ &= \sum_{\varphi \leq_K \psi,\, supp(\varphi) \subseteq E} hg[K](\psi)(\varphi) \\ &= \sum_{\varphi \leq_K \psi \bullet \mathbf{1}_E} hg[K](\psi)(\varphi) \\ &= \frac{\sum_{\varphi \leq_K \psi \bullet \mathbf{1}_E} \binom{\psi \bullet \mathbf{1}_E}{\varphi}}{\binom{L}{K}} \;=\; \frac{\binom{L_E}{K}}{\binom{L}{K}} \qquad \text{by Lemma 1.6.2.} \end{aligned}$$

Then:

$$\begin{aligned} hg[K](\psi)\big|_{\overline{\mathbf{1}_E}^{\bullet}}(\varphi) \;&=\; \sum_{\varphi \leq_K \psi} \frac{hg[K](\psi)(\varphi) \cdot \overline{\mathbf{1}_E}^{\bullet}(\varphi)}{hg[K](\psi) \models \overline{\mathbf{1}_E}^{\bullet}} \,\big|\varphi\big\rangle \\ &= \sum_{\varphi \leq_K \psi \bullet \mathbf{1}_E} \frac{\prod_x \binom{\psi(x)}{\varphi(x)}}{\binom{L}{K}} \cdot \frac{\binom{L}{K}}{\binom{L_E}{K}} \,\big|\varphi\big\rangle \\ &= \sum_{\varphi \leq_K \psi \bullet \mathbf{1}_E} \frac{\binom{\psi \bullet \mathbf{1}_E}{\varphi}}{\binom{L_E}{K}} \,\big|\varphi\big\rangle \;=\; hg[K](\psi \bullet \mathbf{1}_E). \end{aligned}$$

2 Similarly. ❑

We illustrate how to solve a famous riddle via updating a draw distribution.

Example 6.2.3. We start by recalling the ordered draw channel $Ot_- \colon \mathcal{N}_*(X) \nrightarrow \mathcal{L}_*(X) \times \mathcal{N}(X)$ from Figure 3.1. For convenience, we recall that it is given by:

$$Ot_-(\psi) \;=\; \sum_{x \in supp(\psi)} \frac{\psi(x)}{\|\psi\|} \big|[x],\, \psi - 1|x\rangle\big\rangle.$$

We recall that these maps can be iterated, via Kleisli composition $\circ$, as described in (3.18).

Below we like to incorporate the evidence that the last drawn colour is in a subset $U \subseteq X$. We thus extend U to $last(U) \subseteq \mathcal{L}_*(X)$, given by:

$$last(U) = \{ [x_1, \ldots, x_n] \in \mathcal{L}_*(X) \mid x_n \in U \}.$$

We now come to the so-called Monty Hall problem. It is a famous riddle in probability due to [150], see also *e.g.* [61, 158]:

> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

We describe the three doors via a set $D = \{1, 2, 3\}$ and the situation in which to make the first choice as a multiset $\psi = 1|1\rangle + 1|2\rangle + 1|3\rangle$. One may also view ψ as an urn from which to draw numbered balls. Let's assume without loss of generality that the car is behind door 2. We shall represent this logically by using a (sharp) 'goat' predicate $\mathbf{1}_G \colon \{1, 2, 3\} \to [0, 1]$, for $G = \{1, 3\}$.

Your first draw yields a distribution:

$$Ot_-(\psi) = \tfrac{1}{3}\big|[1],\ 1|2\rangle + 1|3\rangle\big\rangle + \tfrac{1}{3}\big|[2],\ 1|1\rangle + 1|3\rangle\big\rangle + \tfrac{1}{3}\big|[3],\ 1|1\rangle + 1|2\rangle\big\rangle.$$

Next, the host's opening of a door with a goat is modeled by a new draw, but we know that a goat appears. The resulting distribution is described by another draw followed by an update:

$$\begin{aligned}&(Ot_- \circ Ot_-)(\psi)\big|_{\mathbf{1}_{last(G)} \otimes \mathbf{1}} \\ &= \tfrac{1}{4}\big|[1,3],1|2\rangle\big\rangle + \tfrac{1}{4}\big|[2,1],1|3\rangle\big\rangle + \tfrac{1}{4}\big|[2,3],1|1\rangle\big\rangle + \tfrac{1}{4}\big|[3,1],1|2\rangle\big\rangle.\end{aligned}$$

Inspection of this outcome shows that for two of your three choices — when you choose 1 or 3, in the first and last term — it makes sense to change your choice, because the car is behind the remaining door 2. Hence changing is better than sticking to your original choice.

We have given a formal account of the situation. More informally, the host knows where the car is, so his choice is not arbitrary. By opening a door with a goat behind it, the host is giving you information that you can exploit to improve your choice: two of your possible choices are wrong, but in those two out three cases the host gives you information how to correct your choice.

We conclude with a well known result, namely that a multinomial distribution can be obtained from conditioning parallel Poisson distributions. This result can be found for instance in [145, §6.4], in binary form. Recall that a

Poisson distribution has a rate/intensity λ as parameter, giving the number of events per time unit. The Poisson distribution then describes the probability of k events, per time unit. When we put n Poisson processes in parallel, with rates $\lambda_1, \ldots, \lambda_n$ and condition on getting K events in total, the resulting probability can also be obtained as a multinomial with K-sized draws, and with state/urn proportional to the rates $\lambda_1, \ldots, \lambda_n$.

Theorem 6.2.4. *Let $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_{\geq 0}$ be given, with $\lambda := \sum_i \lambda_i > 0$. For $K \in \mathbb{N}$, consider the sharp predicate $\mathrm{sum}_K \colon \mathbb{N}^m \to \{0, 1\}$ given my:*

$$sum_K(n_1, \ldots, n_m) = 1 \iff n_1 + \cdots + n_m = K.$$

Then, for $n_1, \ldots, n_m \in \mathbb{N}$ with $\sum_i n_i = K$,

$$\begin{aligned}
&(pois[\lambda_1] \otimes \cdots \otimes pois[\lambda_m])\big|_{sum_K} \\
&= \sum_{n_1, \ldots, n_m,\, \sum_i n_i = K} mn[K]\Big(\textstyle\sum_i \frac{\lambda_i}{\lambda} |i\rangle\Big)\Big(\textstyle\sum_i n_i |i\rangle\Big)\,\big|n_1, \ldots, n_m\big\rangle.
\end{aligned}$$

Proof. We first compute the validity:

$$\begin{aligned}
&pois[\lambda_1] \otimes \cdots \otimes pois[\lambda_m] \models sum_K \\
&= \sum_{n_1, \ldots, n_m,\, \sum_i n_i = K} pois[\lambda_1](n_1) \cdot \ldots \cdot pois[\lambda_m](n_m) \\
&= \sum_{n_1, \ldots, n_m,\, \sum_i n_i = K} e^{-\lambda_1} \cdot \frac{\lambda_1^{n_1}}{n_1!} \cdot \ldots \cdot e^{-\lambda_m} \cdot \frac{\lambda_2^{n_m}}{n_m!} \\
&= \frac{e^{-(\lambda_1 + \cdots + \lambda_m)}}{K!} \sum_{n_1, \ldots, n_m,\, \sum_i n_i = K} \binom{K}{n_1, \ldots, n_m} \cdot \lambda_1^{n_1} \cdot \ldots \cdot \lambda_m^{n_m} \\
&\overset{(1.26)}{=} \frac{e^{-\lambda}}{K!} \cdot \lambda^K \\
&= pois[\lambda](K).
\end{aligned}$$

The update then becomes:

$$\begin{aligned}
&(pois[\lambda_1] \otimes \cdots \otimes pois[\lambda_m])\big|_{sum_K} \\
&= \sum_{n_1,\ldots,n_m,\, \sum_i n_i = K} \frac{pois[\lambda_1](n_1) \cdot \ldots \cdot pois[\lambda_m](n_m)}{pois[\lambda_1] \otimes \cdots \otimes pois[\lambda_m] \models sum_K} \,\big|n_1, \ldots, n_m\big\rangle \\
&= \sum_{n_1,\ldots,n_m,\, \sum_i n_i = K} \frac{e^{-\lambda_1} \cdot \frac{\lambda_1^{n_1}}{n_1!} \cdot \ldots \cdot e^{-\lambda_m} \cdot \frac{\lambda_2^{n_m}}{n_m!}}{\frac{e^{-\lambda}}{K!} \cdot \lambda^K} \,\big|n_1, \ldots, n_m\big\rangle \\
&= \sum_{n_1,\ldots,n_m,\, \sum_i n_i = K} \binom{K}{n_1, \ldots, n_m} \cdot \textstyle\prod_i (\frac{\lambda_i}{\lambda})^{n_i} \,\big|n_1, \ldots, n_m\big\rangle \\
&= \sum_{n_1,\ldots,n_m,\, \sum_i n_i = K} mn[K]\big(\textstyle\sum_i \frac{\lambda_i}{\lambda}|i\rangle\big)\big(\textstyle\sum_i n_i|i\rangle\big)\,\big|n_1, \ldots, n_m\big\rangle. \qquad \blacksquare
\end{aligned}$$

The next result shows how hypergeometric distributions can be expressed in terms of multinomial distributions, where, somewhat remarkably, the state ω is completely irrelevant.

Theorem 6.2.5. *Let $K, L \in \mathbb{N}$ be given with $K \leq L$. Fix $\psi \in \mathcal{N}[L](X)$ and consider the sum of natural multisets as a sharp predicate $\mathrm{sum}_\psi \colon \mathcal{N}[K](X) \times \mathcal{N}[L-K](X) \to \{0,1\}$ given by:*

$$sum_\psi(\varphi, \varphi') = 1 \iff \varphi + \varphi' = \psi.$$

1 For each state $\omega \in \mathcal{D}(X)$,

$$mn[K](\omega) \otimes mn[L-K](\omega) \models sum_\psi \;=\; mn[L](\omega)(\psi).$$

2 Again for each state $\omega \in \mathcal{D}(X)$,

$$hg[K](\psi) \;=\; \big((mn[K](\omega) \otimes mn[L-K](\omega))\big|_{sum_\psi}\big)[1, 0].$$

Proof. 1 By Exercise 3.3.8 we first get:

$$\begin{aligned}
&mn[K](\omega) \otimes mn[L-K](\omega) \models sum_\psi \\
&= \sum_{\varphi \in \mathcal{N}[K](X),\, \varphi' \in \mathcal{N}[L-K](X),\, \varphi + \varphi' = \psi} (mn[K](\omega) \otimes mn[L-K](\omega))(\varphi, \varphi') \\
&= \big(sum \gg\!= (mn[K](\omega) \otimes mn[L-K](\omega))\big)(\psi) \\
&= mn[L](\omega)(\psi).
\end{aligned}$$

2 Next we do the update:

$$\begin{aligned}
&(mn[K](\omega) \otimes mn[L-K](\omega))\big|_{sum_\psi} \\
&= \sum_{\varphi \in \mathcal{N}[K](X),\, \varphi' \in \mathcal{N}[L-K](X),\, \varphi+\varphi'=\psi} \frac{mn[K](\omega)(\varphi) \cdot mn[L-K](\omega)(\varphi')}{mn[K](\omega) \otimes mn[L-K](\omega) \models sum_\psi} \big|\varphi, \varphi'\big\rangle \\
&= \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi} \cdot (\prod_x \omega(x)^{\varphi(x)}) \cdot \binom{\psi - \varphi}{} \cdot (\prod_x \omega(x)^{\psi(x)-\varphi(x)})}{mn[L](\omega)(\psi)} \big|\varphi, \psi - \varphi\big\rangle \\
&= \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi} \cdot \binom{\psi - \varphi}{} \cdot (\prod_x \omega(x)^{\psi(x)})}{\binom{\psi}{} \cdot (\prod_x \omega(x)^{\psi(x)})} \big|\varphi, \psi - \varphi\big\rangle \\
&= \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L}{K}} \big|\varphi, \psi - \varphi\big\rangle, \qquad \text{see Lemma 1.6.4 (1).}
\end{aligned}$$

The first marginal of the latter distribution is the hypergeometric distribution. ❑

Exercises

6.2.1 Prove Theorem 6.2.2 (2) yourself.

6.2.2 Let $K_1, \ldots, K_N \in \mathbb{N}$ with $\psi \in \mathcal{N}[K](X)$ be given, where $K = \sum_i K_i$. Write $sum_\psi \colon \mathcal{N}[K_1](X) \times \cdots \times \mathcal{N}[K_N] \to \{0, 1\}$ for the sharp predicate defined by:

$$sum_\psi(\varphi_1, \ldots, \varphi_N) = 1 \iff \varphi_1 + \cdots + \varphi_N = \psi.$$

Show that, for any $\omega \in \mathcal{D}(X)$,

$$\begin{aligned}
&(mn[K_1](\omega) \otimes \cdots \otimes mn[K_N](\omega))\big|_{sum_\psi} \\
&= \sum_{\varphi_1 \leq_{K_1} \psi, \ldots, \varphi_N \leq_{K_N} \psi,\, \sum_i \varphi_i = \psi} \frac{1}{\binom{\psi}{\varphi_1, \ldots, \varphi_N}} \big|\varphi_1, \ldots, \varphi_N\big\rangle.
\end{aligned}$$

Recall from Exercise 1.6.9 that the inverses of the multinomial coefficients of multisets in the above sum add up to one.

6.3 Forward and backward inference

Forward transformation $\gg=$ of states and backward transformation $=\ll$ of observables can be combined with updating of states. It gives rise to the powerful techniques of forward and backward inference. This will be illustrated in the current section. At the end, these channel-based inference mechanisms are related to crossover inference for joint states.

The next definition captures the two basic patterns — first formulated in this form in [87]. We shall refer to them jointly as channel-based inference, or as reasoning along channels.

Definition 6.3.1. Let $\omega \in \mathcal{D}(X)$ be a state on the domain of a channel $c\colon X \multimap Y$.

1 For a factor $p \in \mathit{Fact}(X)$, we define *forward inference* as transformation along c of the state ω updated with p, as in:

$$c \gg\!= \omega|_p.$$

This is also called *prediction* or *causal reasoning*.

2 For a factor $q \in \mathit{Fact}(Y)$, *backward inference* is updating of ω with the transformed factor:

$$\omega|_{c \mathrel{=\!\ll} q}.$$

This is also called *explanation* or *evidential reasoning*. We shall also refer this operation as *Pearl's update rule*, in contrast with Jeffrey's rule, to be discussed in Section 6.7.

In both cases the distribution ω is often called the *prior distribution* or simply the *prior*. Similarly, $c \gg\!= \omega|_p$ and $\omega|_{c \mathrel{=\!\ll} q}$ are called *posterior distributions* or just *posteriors*.

Thus, with forward inference one first conditions and then performs (forward, state) transformation, whereas for backward inference one first performs (backward, factor) transformation, and then one conditions. We shall illustrate these fundamental inferences mechanisms in several examples. They mostly involve backward inference, since that is the more useful technique. An important first step in these examples is to recognise the channel that is hidden in the description of the problem. It is instructive to try and do this, before reading the analysis and the solution.

Example 6.3.2. We start with the following question from [144, Example 1.12].

Consider two urns. The first contains two white and seven black balls, and the second contains five white and six black balls. We flip a coin and then draw a ball from the first urn or the second urn depending on whether the outcome was heads or tails. What is the conditional probability that the outcome of the toss was heads given that a white ball was selected?

Our analysis involves two sample spaces $\{H, T\}$ for the sides of the coin and

$\{W, B\}$ for the colours of the balls in the urns. The coin distribution is uniform: $\mathit{unif} = \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle$. The above description implicitly contains a channel $c \colon \{H, T\} \to \{W, B\}$, namely:

$$\begin{aligned} c(H) &= \mathit{Flrn}(2|W\rangle + 7|B\rangle) \\ &= \tfrac{2}{9}|W\rangle + \tfrac{7}{9}|B\rangle \end{aligned} \qquad \begin{aligned} c(T) &= \mathit{Flrn}(5|W\rangle + 6|B\rangle) \\ &= \tfrac{5}{11}|W\rangle + \tfrac{6}{11}|B\rangle. \end{aligned}$$

As in the above quote, the first urn is associated with heads and the second one with tails.

The evidence that we have is described in the quote after the word 'given'. It is the point predicate $\mathbf{1}_W$ on the set of colours $\{W, B\}$, indicating that a white ball was selected. This evidence can be pulled back (transformed) along the channel c, to a predicate $c \mathrel{=\!\!\ll} \mathbf{1}_W$ on the sample space $\{H, T\}$. It is given by:

$$(c \mathrel{=\!\!\ll} \mathbf{1}_W)(H) = \sum_{x \in \{W,B\}} c(H)(x) \cdot \mathbf{1}_W(x) = c(H)(W) = \tfrac{2}{9}.$$

Similarly we get $(c \mathrel{=\!\!\ll} \mathbf{1}_W)(T) = c(T)(W) = \frac{5}{11}$.

The answer that we are interested in is obtained by updating the prior *unif* with the transformed evidence $c \mathrel{=\!\!\ll} \mathbf{1}_W$, as given by $\mathit{unif}|_{c \mathrel{=\!\!\ll} \mathbf{1}_H}$. This is an instance of backward inference.

In order to get this answer, we first we compute the validity:

$$\begin{aligned} \mathit{unif} \models c \mathrel{=\!\!\ll} \mathbf{1}_W &= \mathit{unif}(H) \cdot (c \mathrel{=\!\!\ll} \mathbf{1}_W)(H) + \mathit{unif}(T) \cdot (c \mathrel{=\!\!\ll} \mathbf{1}_W)(T) \\ &= \tfrac{1}{2} \cdot \tfrac{2}{9} + \tfrac{1}{2} \cdot \tfrac{5}{11} = \tfrac{1}{9} + \tfrac{5}{22} = \tfrac{67}{198}. \end{aligned}$$

Then:

$$\mathit{unif}|_{c \mathrel{=\!\!\ll} \mathbf{1}_W} = \frac{1/2 \cdot 2/9}{67/198}|H\rangle + \frac{1/2 \cdot 5/11}{67/198}|T\rangle = \tfrac{22}{67}|H\rangle + \tfrac{45}{67}|T\rangle.$$

Thus, the conditional probability of heads is $\frac{22}{67}$. The same outcome is obtained in [144], of course, but there via an application of Bayes' rule.

Example 6.3.3. Consider the following classical question from [159].

A cab was involved in a hit and run accident at night. Two cab companies, Green and Blue, operate in the city. You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue
- A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident, and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved in the accident was Blue rather than Green?

We use as colour set $C = \{G, B\}$ for Green and Blue. There is a prior 'base rate' distribution $\omega = \frac{17}{20}|G\rangle + \frac{3}{20}|B\rangle \in \mathcal{D}(C)$, as in the first bullet above. The reliability information in the second bullet translates into a 'correctness' channel $c\colon \{G, B\} \nrightarrow \{G, B\}$ given by:

$$c(G) = \tfrac{4}{5}|G\rangle + \tfrac{1}{5}|B\rangle \qquad\qquad c(B) = \tfrac{1}{5}|G\rangle + \tfrac{4}{5}|B\rangle.$$

The second bullet also gives evidence of a Blue car. It translates into a point predicate $\mathbf{1}_B$ on $\{G, B\}$. It can be used for backward inference, giving the answer to the query, as posterior:

$$\omega|_{c \mathrel{=\!\!\ll} \mathbf{1}_B} = \tfrac{17}{29}|G\rangle + \tfrac{12}{29}|B\rangle \approx 0.5862|G\rangle + 0.4138|B\rangle.$$

Thus the probability that the Blue car was actually involved in the incident is a bit more that 41%. This may seem like a relatively low probability, given that the evidence says 'Blue taxicab' and that observations are 80% accurate. But this low percentage is explained by the fact that are relatively few Bleu taxicabes in the first place, namely only 15%. This is in the prior, base rate distribution ω. It is argued in [159] that humans find it difficult to take such base rates (or priors) into account. They call this phenomenon "base rate neglect", see also [57].

Notice that the channel c is used to accomodate the uncertainty of observations: the sharp point observation 'Blue taxicab' is transformed into a fuzzy predicate $c \mathrel{=\!\!\ll} \mathbf{1}_B$. The latter is $G \mapsto \frac{1}{5}$ and $B \mapsto \frac{4}{5}$. Updating ω with this predicate gives the claimed outcome.

Example 6.3.4. We continue in the setting of Example 2.2.2, with a teacher in a certain mood, and predictions about pupils' performances depending on the mood. We now assume that the pupils have done rather poorly, with no-one scoring above 5, as described by the following evidence/predicate q on the set of grades $Y = \{1, 2, \ldots, 10\}$.

$$q = \tfrac{1}{10} \cdot \mathbf{1}_1 + \tfrac{3}{10} \cdot \mathbf{1}_2 + \tfrac{3}{10} \cdot \mathbf{1}_3 + \tfrac{2}{10} \cdot \mathbf{1}_4 + \tfrac{1}{10} \cdot \mathbf{1}_5.$$

The validity of this predicate q in the predicted state $c \gg\!= \sigma$ is:

$$c \gg\!= \sigma \models q \;=\; \sigma \models c \mathrel{=\!\!\ll} q \;=\; \tfrac{299}{4000} = 0.07475.$$

The interested reader may wish to check that the updated state $\sigma' = \sigma|_{c \mathrel{=\!\!\ll} q}$ via backward inference is:

$$\sigma' = \tfrac{77}{299}|p\rangle + \tfrac{162}{299}|n\rangle + \tfrac{60}{299}|o\rangle \approx 0.2575|p\rangle + 0.5418|n\rangle + 0.2007|o\rangle.$$

Interestingly, after updating the teacher has more realistic view in the sense that the validity of the predicate q has risen to $c \gg\!= \sigma' \models q = \frac{15577}{149500} \approx 0.1042$. This is one way how the mind can adapt to external evidence.

Example 6.3.5. Recall the Medicine-Blood Table (1.18) with data on different types of medicine via a set $M = \{0, 1, 2\}$ and blood pressure via the set $B = \{H, L\}$. From the table we can extract a channel $b\colon M \nrightarrow B$ describing the blood pressure distribution for each medicine type. This channel is obtained by column-wise frequentist learning:

$$b(0) = \tfrac{2}{3}|H\rangle + \tfrac{1}{3}|L\rangle \qquad b(1) = \tfrac{7}{9}|H\rangle + \tfrac{2}{9}|L\rangle \qquad b(2) = \tfrac{5}{8}|H\rangle + \tfrac{3}{8}|L\rangle.$$

The prior medicine distribution $\omega = \frac{3}{20}|0\rangle + \frac{9}{20}|1\rangle + \frac{2}{5}|2\rangle$ is obtained from the totals row in the table.

The predicted state $b \gg\!= \omega$ is $\frac{7}{10}|H\rangle + \frac{3}{10}|L\rangle$. It is the distribution that is learnt from the totals column in Table (1.18). Suppose we wish to focus on the people that take either medicine 1 or 2. We do so by conditioning, via the subset $E = \{1, 2\} \subseteq B$, with associated sharp predicate $\mathbf{1}_E\colon B \to [0, 1]$. Then:

$$\omega \models \mathbf{1}_E = \tfrac{9}{20} + \tfrac{2}{5} = \tfrac{17}{20} \quad \text{so} \quad \omega|_{\mathbf{1}_E} = \frac{9/20}{17/20}|1\rangle + \frac{2/5}{17/20}|2\rangle = \tfrac{9}{17}|1\rangle + \tfrac{8}{17}|2\rangle.$$

Forward reasoning, precisely as in Definition 6.3.1 (1), gives:

$$\begin{aligned} b \gg\!= (\omega|_{\mathbf{1}_E}) &= b \gg\!= (\tfrac{9}{17}|1\rangle + \tfrac{8}{17}|2\rangle) \\ &= (\tfrac{9}{17} \cdot \tfrac{7}{9} + \tfrac{8}{17} \cdot \tfrac{5}{8})|H\rangle + (\tfrac{9}{17} \cdot \tfrac{2}{9} + \tfrac{8}{17} \cdot \tfrac{3}{8})|H\rangle \\ &= \tfrac{12}{17}|H\rangle + \tfrac{5}{17}|L\rangle. \end{aligned}$$

This shows the distribution of high and low blood pressure among people using medicine 1 or 2.

We turn to backward reasoning. Suppose that we have evidence $\mathbf{1}_H$ on $\{H, L\}$ of high blood pressure. What is then the associated distribution of medicine usage? It is obtained in several steps:

$$\begin{aligned} (b \mathrel{=\!\!\ll} \mathbf{1}_H)(x) &= \sum\nolimits_y b(x)(y) \cdot \mathbf{1}_H(y) = b(x)(H) \\ \omega \models b \mathrel{=\!\!\ll} \mathbf{1}_H &= \sum\nolimits_x \omega(x) \cdot (b \mathrel{=\!\!\ll} \mathbf{1}_H)(x) = \sum\nolimits_x \omega(x) \cdot b(x)(H) \\ &= \tfrac{3}{20} \cdot \tfrac{2}{3} + \tfrac{9}{20} \cdot \tfrac{7}{9} + \tfrac{2}{5} \cdot \tfrac{5}{8} = \tfrac{7}{10} \\ \omega|_{b \mathrel{=\!\!\ll} \mathbf{1}_H} &= \sum\nolimits_x \frac{\omega(x) \cdot (b \mathrel{=\!\!\ll} \mathbf{1}_H)(x)}{\omega \models b \mathrel{=\!\!\ll} \mathbf{1}_H} \big|x\rangle \\ &= \frac{3/20 \cdot 2/3}{7/10}|0\rangle + \frac{9/20 \cdot 7/9}{7/10}|1\rangle + \frac{2/5 \cdot 5/8}{7/10}|2\rangle \\ &= \tfrac{1}{7}|0\rangle + \tfrac{1}{2}|1\rangle + \tfrac{5}{14}|2\rangle \approx 0.1429|0\rangle + 0.5|1\rangle + 0.3571|2\rangle. \end{aligned}$$

We can also reason with 'soft' evidence, using the full power of fuzzy predicates. Suppose we are only 95% sure that the blood pressure is high, due to some measurement uncertainty. Then we can use as evidence the predicate

$q\colon B \to [0,1]$ given by $q(H) = 0.95$ and $q(L) = 0.05$. It yields:

$$\omega|_{b \prec\!\ll q} \approx 0.1434|0\rangle + 0.4963|1\rangle + 0.3603|2\rangle.$$

We see a slight difference wrt. the outcome with sharp evidence.

Example 6.3.6. The following question comes from [158, §6.1.3] (and is also used in [129]).

One fish is contained within the confines of an opaque fishbowl. The fish is equally likely to be a piranha or a goldfish. A sushi lover throws a piranha into the fish bowl alongside the other fish. Then, immediately, before either fish can devour the other, one of the fish is blindly removed from the fishbowl. The fish that has been removed from the bowl turns out to be a piranha. What is the probability that the fish that was originally in the bowl by itself was a piranha?

Let's use the letters 'p' and 'g' for piranha and goldfish. We are looking at a situation with multiple fish in a bowl, where we cannot distinguish the order. Hence we describe the contents of the bowl as a (natural) muliset over $\{p, g\}$, that is, as an element of $\mathcal{N}(\{p, g\})$. The initial situation can then be described as a distribution $\omega \in \mathcal{D}(\mathcal{N}(\{p, g\}))$ with:

$$\omega = \tfrac{1}{2}\big|\, 1|p\rangle \,\big\rangle + \tfrac{1}{2}\big|\, 1|g\rangle \,\big\rangle.$$

Adding a piranha to the bowl involves a function $A\colon \mathcal{N}(\{p, g\}) \to \mathcal{N}(\{p, g\})$, such that $A(\sigma) = (\sigma(p) + 1)|p\rangle + \sigma(g)|g\rangle$. It forms a deterministic channel.

We use a piranha predicate $P\colon \mathcal{N}(\{p, g\}) \to [0, 1]$ that gives the likelihood $P(\sigma)$ of taking a piranha from a multiset/bowl σ. Thus:

$$P(\sigma) \coloneqq \mathit{Flrn}(\sigma)(p) \overset{(2.13)}{=} \frac{\sigma(p)}{\sigma(p) + \sigma(g)}.$$

We have now collected all ingredients to answer the question via backward inference along the deterministic channel A. It involves the following steps.

$$\begin{aligned}
(A \mathrel{-\!\!\ll} P)(\sigma) &= P(A(\sigma)) = \frac{\sigma(p) + 1}{\sigma(p) + 1 + \sigma(g)} \\
\omega \models A \mathrel{-\!\!\ll} P &= \tfrac{1}{2} \cdot (A \mathrel{-\!\!\ll} P)(1|p\rangle) + \tfrac{1}{2} \cdot (A \mathrel{-\!\!\ll} P)(1|g\rangle) \\
&= \tfrac{1}{2} \cdot \tfrac{1+1}{1+1+0} + \tfrac{1}{2} \cdot \tfrac{0+1}{0+1+1} = \tfrac{1}{2} + \tfrac{1}{4} = \tfrac{3}{4} \\
\omega|_{A \mathrel{-\!\!\ll} P} &= \frac{1/2 \cdot 1}{3/4}\big|\, 1|p\rangle \,\big\rangle + \frac{1/2 \cdot 1/2}{3/4}\big|\, 1|g\rangle \,\big\rangle \\
&= \tfrac{2}{3}\big|\, 1|p\rangle \,\big\rangle + \tfrac{1}{3}\big|\, 1|g\rangle \,\big\rangle.
\end{aligned}$$

Hence the answer to the question in the beginning of this example is: $\frac{2}{3}$ probability that the original fish is a piranha.

Example 6.3.7. In [146, §20.1] a situation is described with five different bags, numbered $1, \ldots, 5$, each containing its own mixture of cherry (C) and lime (L) candies. This situation can be described via a candy channel:

$$B \longrightarrow\!\!\!\!\circ\; \{C, L\} \quad \text{where} \quad B = \{1, 2, 3, 4, 5\} \quad \text{and} \quad \begin{cases} c(1) = 1|C\rangle \\ c(2) = \frac{3}{4}|C\rangle + \frac{1}{4}|L\rangle \\ c(3) = \frac{1}{2}|C\rangle + \frac{1}{2}|L\rangle \\ c(4) = \frac{1}{4}|C\rangle + \frac{3}{4}|L\rangle \\ c(5) = 1|L\rangle. \end{cases}$$

The initial bag distribution is $\omega = \frac{1}{10}|1\rangle + \frac{1}{5}|2\rangle + \frac{2}{5}|3\rangle + \frac{1}{5}|4\rangle + \frac{1}{10}|5\rangle$.

In the situation described in [146, §20.1] the sample space of bags B is regarded as hidden (not directly observable), in a scenario where a new bag $i \in B$ is given and candies are drawn from it, inspected and returned. It turns out that 10 consecutive draws yield a lime candy; what can we then infer about the distribution of bags?

Transforming the lime point predicate $\mathbf{1}_L$ along channel c yields the fuzzy predicate $c \mathrel{=\!\!\ll} \mathbf{1}_L \colon B \to [0, 1]$ given by:

$$c \mathrel{=\!\!\ll} \mathbf{1}_L = \bigotimes_i c(i)(L) \cdot \mathbf{1}_i = \tfrac{1}{4} \cdot \mathbf{1}_2 \otimes \tfrac{1}{2} \cdot \mathbf{1}_3 \otimes \tfrac{3}{4} \cdot \mathbf{1}_4 \otimes 1 \cdot \mathbf{1}_5.$$

The question is what we learn about the bag distribution after observing this predicate 10 consecutive times? This involves computing:

$$\begin{aligned} \omega|_{c \mathrel{=\!\!\ll} \mathbf{1}_L} &= \tfrac{1}{10}|2\rangle + \tfrac{2}{5}|3\rangle + \tfrac{3}{10}|4\rangle + \tfrac{1}{5}|5\rangle \\ \omega|_{c \mathrel{=\!\!\ll} \mathbf{1}_L}|_{c \mathrel{=\!\!\ll} \mathbf{1}_L} &= \omega|_{(c \mathrel{=\!\!\ll} \mathbf{1}_L)\&(c \mathrel{=\!\!\ll} \mathbf{1}_L)} = \omega|_{(c \mathrel{=\!\!\ll} \mathbf{1}_L)^2} \\ &= \tfrac{1}{26}|2\rangle + \tfrac{4}{13}|3\rangle + \tfrac{9}{26}|4\rangle + \tfrac{4}{13}|5\rangle \\ &\approx 0.0385|2\rangle + 0.308|3\rangle + 0.346|4\rangle + 0.308|5\rangle \\ \omega|_{c \mathrel{=\!\!\ll} \mathbf{1}_L}|_{c \mathrel{=\!\!\ll} \mathbf{1}_L}|_{c \mathrel{=\!\!\ll} \mathbf{1}_L} &= \omega|_{(c \mathrel{=\!\!\ll} \mathbf{1}_L)\&(c \mathrel{=\!\!\ll} \mathbf{1}_L)\&(c \mathrel{=\!\!\ll} \mathbf{1}_L)} = \omega|_{(c \mathrel{=\!\!\ll} \mathbf{1}_L)^3} \\ &= \tfrac{1}{76}|2\rangle + \tfrac{4}{19}|3\rangle + \tfrac{27}{76}|4\rangle + \tfrac{8}{19}|5\rangle \\ &\approx 0.0132|2\rangle + 0.211|3\rangle + 0.355|4\rangle + 0.421|5\rangle \quad \ldots \end{aligned}$$

Figure 20.1 in [146] gives a plot of these distributions; it is reconstructed here in Figure 6.1 via the above formulas. It shows that bag 5 quickly becomes most likely — as expected since it contains most lime candies — and that bag 1 is impossible after drawing the first lime.

Example 6.3.8. Medical tests form standard examples of Bayesian reasoning, via backward inference, see *e.g.* Exercises 6.3.1 and 6.3.2 below. Here we look at Covid-19 ("Corona") which is interesting because its standard PCR-test has

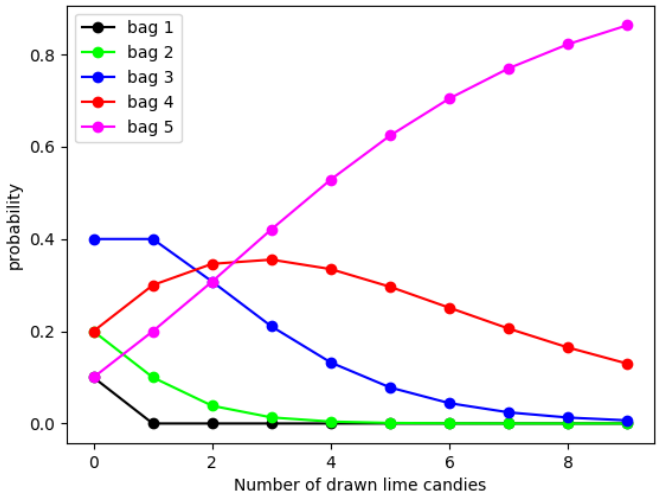


Figure 6.1 Posterior, updated bag distributions $\omega|_{(c \ll \mathbf{1}_L)^n}$ for $n = 0, 1, \ldots, 10$, aligned vertically, after n candy draws that all happen to be lime.

low false positives but high false negatives, and moreover these false negatives depend on the day after infection.

The Covid-19 PCR-test has almost no false positives. This means that if you do not have the disease, then the likelihood of a (false) postive test is very low. In our calculations below we put it at 1%, independently of the day that you get tested. In contrast, the PCR-test has considerable false negative rates, which depend on the day after infection. The plot at the top in Figure 6.2 gives an indication; it does not precisely reflect the medical reality, but it provides a reasonable approximation, good enough for our calculation. This plot shows that if you are infected at day 0, then a test at this day or the day after (day 1) will surely be negative. On the second day after your infection a PCR-test might start to detect, but still there is only a 20% chance of a positive outcome. This probability increases and after on day 6 the likelihood of a positive test has risen to 80%.

How to formalise this situation? We use the following three sample spaces, for Covid (C), days after infection (D), and test outcome (T).

$$C = \{c, c^\perp\} \qquad D = \{0, 1, 2, 3, 4, 5, 6\} \qquad T = \{p, n\}.$$

The test probabilities are then captured via a test channel $t \colon C \times D \nrightarrow T$, given

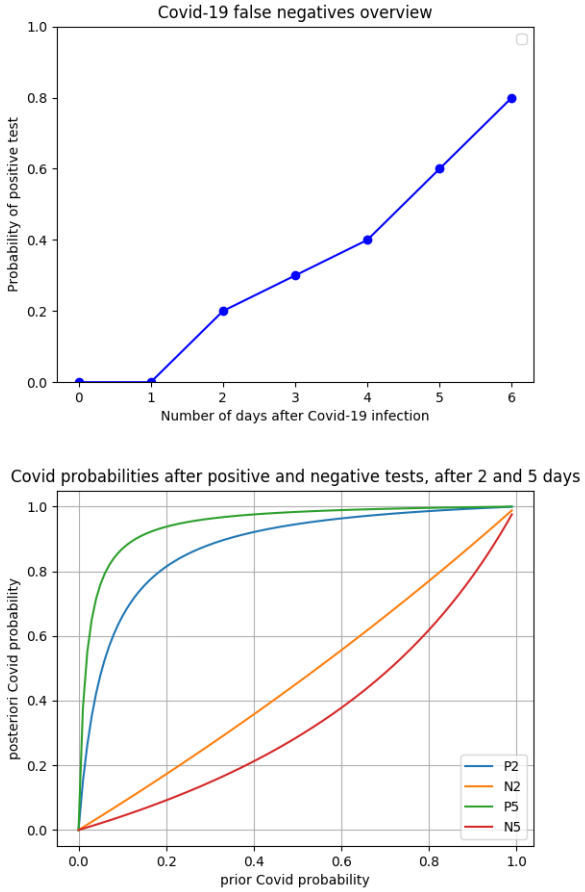


Figure 6.2 Covid-19 false negatives and posteriors after tests. In the lower plot P2 means: positive test after 2 days, that is, in state $(r|c\rangle + (1-r)|c^{\perp}\rangle) \otimes \varphi_2$, where $r \in [0,1]$ is the prior Covid probability. Similarly for N2, P5, N5.

in the following way. The first equation captures the false positives, and the

second one the false negatives, as in the plot at the top of Figure 6.2.

$$t(c^\perp, i) \;=\; \tfrac{1}{100}|p\rangle + \tfrac{99}{100}|n\rangle$$

$$t(c, i) \;=\; \begin{cases} 1|n\rangle & \text{if } i = 0 \text{ or } i = 1 \\ \frac{2}{10}|p\rangle + \frac{8}{10}|n\rangle & \text{if } i = 2 \\ \frac{3}{10}|p\rangle + \frac{7}{10}|n\rangle & \text{if } i = 3 \\ \frac{4}{10}|p\rangle + \frac{6}{10}|n\rangle & \text{if } i = 4 \\ \frac{6}{10}|p\rangle + \frac{4}{10}|n\rangle & \text{if } i = 5 \\ \frac{8}{10}|p\rangle + \frac{2}{10}|n\rangle & \text{if } i = 6 \end{cases}$$

In practice it is often difficult to determine the precise date of inffection. We shall consider two cases, where the infection happened (around) two and five days ago, via the two distributions:

$$\varphi_2 \;=\; \tfrac{1}{4}|1\rangle + \tfrac{1}{2}|2\rangle + \tfrac{1}{4}|3\rangle \qquad\qquad \varphi_5 \;=\; \tfrac{1}{4}|4\rangle + \tfrac{1}{2}|5\rangle + \tfrac{1}{4}|6\rangle.$$

Let's consider the case where we have no (prior) information about the likelihood that the person that is going to be tested has the disease. Therefore we use as prior $\omega = \mathit{unif}_C = \frac{1}{2}|c\rangle + \frac{1}{2}|c^\perp\rangle$.

Suppose in this situation we have a positive test, say after two days. What do we then learn about the disease probability? Our evidence is the postive test predicate $\mathbf{1}_p$ on T, which can be transformed to $t \mathrel{=\!\!\ll} \mathbf{1}_p$ on $C \times D$. We can use this predicate to update the joint state $\omega \otimes \varphi_2$. The distribution that we are after is the first marginal of this updated state, as in:

$$\Big((\omega \otimes \varphi_2)\big|_{t \mathrel{=\!\!\ll} \mathbf{1}_p}\Big)[1, 0] \;\in\; \mathcal{D}(C).$$

We shall go through the computation step-by-step. First,

$$\begin{aligned} t \mathrel{=\!\!\ll} \mathbf{1}_p \;&=\; \sum_{x \in C, y \in D} t(x, y)(p) \cdot \mathbf{1}_{(x,y)} \\ &= \tfrac{2}{10} \cdot \mathbf{1}_{(c,2)} + \tfrac{3}{10} \cdot \mathbf{1}_{(c,3)} + \tfrac{4}{10} \cdot \mathbf{1}_{(c,4)} + \tfrac{6}{10} \cdot \mathbf{1}_{(c,5)} + \tfrac{8}{10} \cdot \mathbf{1}_{(c,6)} \\ &\quad + \tfrac{1}{100} \cdot \mathbf{1}_{(c^\perp,0)} + \tfrac{1}{100} \cdot \mathbf{1}_{(c^\perp,1)} + \tfrac{1}{100} \cdot \mathbf{1}_{(c^\perp,2)} + \tfrac{1}{100} \cdot \mathbf{1}_{(c^\perp,3)} \\ &\quad + \tfrac{1}{100} \cdot \mathbf{1}_{(c^\perp,4)} + \tfrac{1}{100} \cdot \mathbf{1}_{(c^\perp,5)} + \tfrac{1}{100} \cdot \mathbf{1}_{(c^\perp,6)}. \end{aligned}$$

Then:

$$\begin{aligned} \omega \otimes \varphi_2 \models t \mathrel{=\!\!\ll} \mathbf{1}_p \;&=\; \sum_{x \in C, y \in D} \omega(x) \cdot \varphi_2(y) \cdot (t \mathrel{=\!\!\ll} \mathbf{1}_p)(x, y) \\ &= \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{2}{10} + \tfrac{1}{2} \cdot \tfrac{1}{4} \cdot \tfrac{3}{10} + \tfrac{1}{2} \cdot \tfrac{1}{4} \cdot \tfrac{1}{100} + \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{100} + \tfrac{1}{2} \cdot \tfrac{1}{4} \cdot \tfrac{1}{100} \\ &= \tfrac{40+30+1+2+1}{800} \;=\; \tfrac{37}{400}. \end{aligned}$$

But then:

$$\begin{aligned}(\omega \otimes \varphi_2)\big|_{t \preccurlyeq \mathbf{1}_p} &= \tfrac{1/2 \cdot 1/2 \cdot 2/10}{37/400}\big|c, 2\rangle + \tfrac{1/2 \cdot 1/4 \cdot 3/10}{37/400}\big|c, 3\rangle + \tfrac{1/2 \cdot 1/4 \cdot 1/100}{37/400}\big|c^\perp, 1\rangle \\ &\quad + \tfrac{1/2 \cdot 1/2 \cdot 1/100}{37/400}\big|c^\perp, 2\rangle + \tfrac{1/2 \cdot 1/4 \cdot 1/100}{37/400}\big|c^\perp, 3\rangle \\ &= \tfrac{20}{37}\big|c, 2\rangle + \tfrac{15}{37}\big|c, 3\rangle + \tfrac{1}{74}\big|c^\perp, 1\rangle + \tfrac{1}{37}\big|c^\perp, 2\rangle + \tfrac{1}{74}\big|c^\perp, 3\rangle.\end{aligned}$$

Finally:

$$\Big((\omega \otimes \varphi_2)\big|_{t \preccurlyeq \mathbf{1}_p}\Big)[1, 0] = \tfrac{35}{37}|c\rangle + \tfrac{2}{37}|c^\perp\rangle \approx 0.946|c\rangle + 0.054|c^\perp\rangle.$$

Hence a positive test changes the a priori likelihood of 50% to about 95%. In a similar way one can compute that:

$$\Big((\omega \otimes \varphi_2)\big|_{t \preccurlyeq \mathbf{1}_n}\Big)[1, 0] = \tfrac{165}{363}|c\rangle + \tfrac{198}{363}|c^\perp\rangle \approx 0.455|c\rangle + 0.545|c^\perp\rangle.$$

We see that a negative test, 2 days after infection, reduces the prior disease probability of 50% only slightly, namely to 45%.

Doing the test around 5 days after infection gives a bit more certainty, especially in the case of a negative test:

$$\begin{aligned}\Big((\omega \otimes \varphi_5)\big|_{t \preccurlyeq \mathbf{1}_p}\Big)[1, 0] &= \tfrac{60}{61}|c\rangle + \tfrac{1}{61}|c^\perp\rangle \approx 0.984|c\rangle + 0.016|c^\perp\rangle \\ \Big((\omega \otimes \varphi_5)\big|_{t \preccurlyeq \mathbf{1}_n}\Big)[1, 0] &= \tfrac{40}{139}|c\rangle + \tfrac{99}{139}|c^\perp\rangle \approx 0.288|c\rangle + 0.712|c^\perp\rangle.\end{aligned}$$

The lower plot in Figure 6.2 gives a more elaborate description, for different disease probabilities r in a prior $\omega = r|c\rangle + (1 - r)|c^\perp\rangle$ as used above. We see that a postive test outcome quickly gives certainty about having the disease. But a negative test outcome gives only a little bit information with respect to the prior — which in this plot can be represented as the diagonal. For this reason, if you get a negative PCR-test, often a second test is done a few days later.

In the next two examples we look at estimating the number of fish in a pond by counting marked fishes, first in multinomial mode and then in hypergeometric mode.

Example 6.3.9. Capture and recapture is a methodology used in ecology to estimate the size of a population. So imagine we are looking at a pond and we wish to learn the number of fish. We catch twenty of them, mark them, and throw them back. Subsequently we catch another twenty, and find out that five of them are marked. What do we learn about the number of fish?

The number of fish in the pond must be at least 20. Let's assume the maximal number is 300. We will be considering units of 10 fish. Hence the underlying

sample space F together with the uniform 'prior' state unif_F is:

$$F = \{20, 30, 40, \ldots, 300\} \qquad \text{with} \qquad \mathit{unif}_F = \textstyle\sum_{x \in F} \frac{1}{29} |\, x \rangle.$$

We now assume that $K = 20$ of the fish in the pond are marked. We can then compute for each value $20, 30, 40, \ldots$ in the fish space F the probability of finding 5 marked fish when 20 of them are caught. In order not to complicate the calculations too much, we catch these 20 fish one by one, check if they are marked, and then throw them back. This means that the probability of catching a marked fish remains the same, and is described by a binomial distribution, see Example 2.1.1 (2). Its parameters are $K = 20$ with probability $\frac{20}{i}$ of catching a marked fish, where $i \in F$ is the assumed total number of fish. This is incorporated in the following 'catch' channel $c \colon F \nrightarrow K+1 = \{0, 1, \ldots, K\}$.

$$c(i) \coloneqq \mathit{bn}[K]\left(\tfrac{K}{i}\right) = \sum_{k \in K+1} \binom{K}{k} \cdot \left(\tfrac{K}{i}\right)^k \cdot \left(\tfrac{i-K}{i}\right)^{K-k} \big|\, k \rangle.$$

Once this is set up, we construct a posterior state by updating the prior with the information that five marked fish have been found. The latter is expressed as point predicate $\mathbf{1}_5 \in \mathit{Pred}(K+1)$ on the codomain of the channel c. We can then do backward inference, as in Definition 6.3.1 (2), and obtain the updated uniform distribution:

$$\mathit{unif}_F\big|_{c \gg \mathbf{1}_5} = \sum_{i \in F} \frac{(20/i)^5 \cdot (i-20/i)^{15}}{\sum_j (20/j)^5 \cdot (j-20/j)^{15}} \big|\, i \rangle.$$

The bar chart of this posterior is in Figure 6.3; it indicates the likelihoods of the various numbers of fish in the pond. One can also compute the expected value (mean) of this posterior; it's 116.5 fish. In case we had caught 10 marked fish out of 20, the expected number would be 47.5.

Note that taking a uniform prior corresponds to the idea that we have no idea about the number of fish in the pond. But possibly we already had a good estimate from previous years. Then we could have used such an estimate as prior distribution, and updated it with this year's evidence.

Example 6.3.10. We take another look at the previous example. There we used the multinomial distribution (in binomial form) for the probability of catching five marked fish, per pond size. This multinomial mode is appropriate for drawing with replacement, which corresponds to returning each fish that we catch to the pond. This is probably not what happens in practice. So let's try to redescribe the capture-recapture model in hypergeometric mode (like in [145, §4.8.3, Ex. 8h]).

Let's write $M = \{m, m^\perp\}$ for the space with elements m for marked and $m^\perp$ for

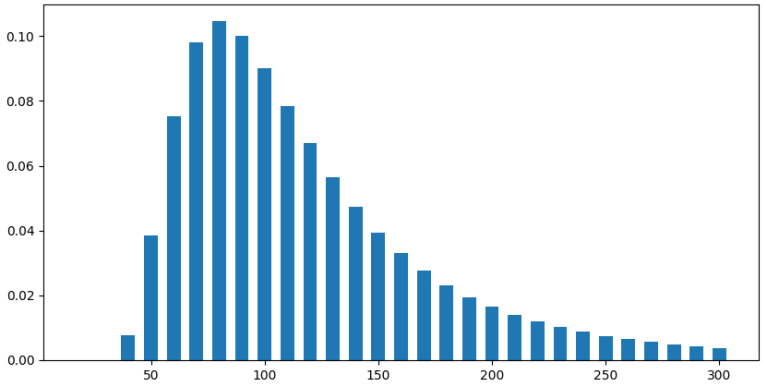


Figure 6.3 The posterior fish number distribution after catching 5 marked fish, in multinomial mode, see Example 6.3.9.

unmarked. Our recapture catch (draw) of $K = 20$ fish, with 5 of them marked, is thus a multiset $\kappa = 5|m\rangle + 15|m^{\perp}\rangle$. The urn from which we draw is the pond, in which the total number of fish is unknown, but we do know that 20 of them are marked. The urn/pond is thus a multiset $20|m\rangle + (i-20)|m^{\perp}\rangle$ with $i \in F$.

We now use a channel:

$$\{20, 30, \ldots, 300\} = F \xrightarrow{\ d\ } \mathcal{N}[K](M) \cong \{0, 1, \ldots, K+1\},$$

given by:

$$d(i) \coloneqq hg[K]\Big(20|m\rangle + (i-20)|m^{\perp}\rangle\Big).$$

The updated pond distribution is then:

$$unif_F\big|_{d \gg \mathbf{1}_\kappa} = \sum_{i \in F} \frac{\binom{i-20}{15}/\binom{i}{20}}{\sum_j \binom{j-20}{15}/\binom{j}{20}} \big|i\rangle.$$

Its bar-chart is in Figure 6.4. It differs minimally from the multinomial one in Figure 6.3. In the hypergeometric case the expected value is 113 fish, against 116.5 in the multinomial case. When the recapture involves 10 marked fish, the expected values are 45.9, against 47.5. As we have already seen in Remark 3.4.3, the hypergeometric distribution on small draws from a large urn look very much like a multinomial distribution.

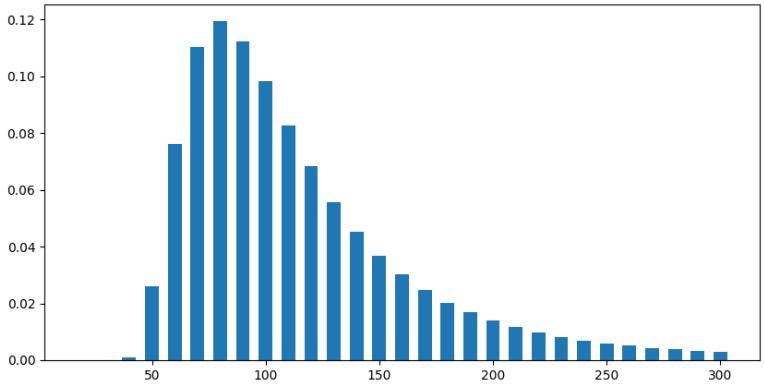


Figure 6.4 The posterior fish number distribution after catching 5 marked fish, in hypergeometric mode, see Example 6.3.10.

6.3.1 Conditioning after state transformation

After all these examples we develop some general results. We start with a fundamental result about conditioning of a transformed state. It can be reformulated as a combination of forward and backward reasoning. This has many consequences.

Theorem 6.3.11. *Let $c\colon X \nrightarrow Y$ be a channel with a state $\omega \in \mathcal{D}(X)$ on its domain and a factor $q \in \mathsf{Fact}(Y)$ on its codomain. Then:*

$$(c \gg\!= \omega)|_q \;=\; c|_q \gg\!= (\omega|_{c \ll q}). \tag{6.2}$$

Proof. For each $y \in Y$,

$$\begin{aligned}
(c \gg\!= \omega)|_q(y) &= \frac{(c \gg\!= \omega)(y) \cdot q(y)}{c \gg\!= \omega \models q} \\
&= \sum_x \frac{\omega(x) \cdot c(x)(y) \cdot q(y)}{\omega \models c =\!\ll q} \\
&= \sum_x \frac{\omega(x) \cdot (c =\!\ll q)(x)}{\omega \models c =\!\ll q} \cdot \frac{c(x)(y) \cdot q(y)}{(c =\!\ll q)(x)} \\
&= \sum_x \omega|_{c =\!\ll q}(x) \cdot \frac{c(x)(y) \cdot q(y)}{c(x) \models q} \\
&= \sum_x \omega|_{c =\!\ll q}(x) \cdot c(x)|_q(y) \\
&= \sum_x \omega|_{c =\!\ll q}(x) \cdot c|_q(x)(y) \\
&= (c|_q \gg\!= (\omega|_{c =\!\ll q}))(y). \qquad \blacksquare
\end{aligned}$$

This result has a number of useful consequences.

Corollary 6.3.12. *For appropriately typed channels, states, and factors:*

1 $(d \circ c)|_q = d|_q \circ c|_{d =\!\ll q}$;

2 $(\langle c, d\rangle \gg\!= \omega)|_{p\otimes q} = \langle c|_p, d|_q\rangle \gg\!= \omega|_{(c =\!\ll p)\&(d =\!\ll q)}$;

3 $((e \otimes f) \gg\!= \omega)|_{p\otimes q} = (e|_p \otimes f|_q) \gg\!= \omega|_{(e =\!\ll p)\otimes(f =\!\ll q)}$.

Proof. 1 $$\begin{aligned}
(d \circ c)|_q(x) = (d \circ c)(x)|_q &\overset{(6.2)}{=} (d \gg\!= c(x))|_q \\
&= d|_q \gg\!= (c(x)|_{d =\!\ll q}) \\
&= d|_q \gg\!= (c|_{d =\!\ll q}(x)) = (d|_q \circ c|_{d =\!\ll q})(x).
\end{aligned}$$

2 $$\begin{aligned}
&(\langle c, d\rangle \gg\!= \omega)|_{p\otimes q} \\
&\overset{(6.2)}{=} (\langle c, d\rangle|_{p\otimes q} \gg\!= (\omega|_{\langle c,d\rangle =\!\ll (p\otimes q)}) \\
&= \langle c|_p, d|_q\rangle \gg\!= (\omega|_{(c =\!\ll p)\&(d =\!\ll q)}) \qquad \text{by Exercises 6.1.11 and 4.3.8.}
\end{aligned}$$

3 Using the same exercises we also get:

$$\begin{aligned}
((e \otimes f) \gg\!= \omega)|_{p\otimes q} &\overset{(6.2)}{=} ((e \otimes f)|_{p\otimes q} \gg\!= (\omega|_{(e\otimes f) =\!\ll (p\otimes q)} \\
&= (e|_p \otimes f|_q) \gg\!= \omega|_{(e =\!\ll p)\otimes(f =\!\ll q)} \qquad \blacksquare
\end{aligned}$$

Earlier we have seen 'crossover update' for a joint state, whereby one updates in one component, and then marginalises in the other, see for instance Example 6.1.2 (2). We shall do this below for joint states $gr(\sigma, c) = \langle id, c\rangle \gg\!= \sigma$ that arise as graph, see Definition 2.5.6, via a channel. The result below appeared in [21], see also [87, 88]. It says that crossover updates on joint graph states can also be done via forward and backward inferences. This is also a consequence of Theorem 6.3.11.

Corollary 6.3.13. *Let $c\colon X \nrightarrow Y$ be a channel with a state $\sigma \in \mathcal{D}(X)$ on its domain. Let $\omega \coloneqq \langle id, c\rangle \gg\!= \sigma \in \mathcal{D}(X \times Y)$ be the associated joint state. Then:*

1 for a factor p on X,

$$\begin{aligned} c \gg\!= (\sigma|_p) &= \omega|_{p\otimes \mathbf{1}}[0,1] \\ &= \pi_2 \gg\!= (\omega|_{\pi_1 \mathrel{\overline{\ll}} p}). \end{aligned}$$

2 for a factor q on Y,

$$\begin{aligned} \sigma|_{c \mathrel{\overline{\ll}} q} &= \omega|_{\mathbf{1}\otimes q}[1,0] \\ &= \pi_1 \gg\!= (\omega|_{\pi_2 \mathrel{\overline{\ll}} q}). \end{aligned}$$

The formulation with the projections π_1, π_2 will be generalised to an inference query in Section 7.10.

Proof. Since:

$$\begin{aligned} &(\langle id, c\rangle \gg\!= \sigma)|_{p\otimes \mathbf{1}}[0,1] \\ &\overset{(6.2)}{=} \pi_2 \gg\!= \Big(\langle id, c\rangle|_{p\otimes \mathbf{1}} \gg\!= (\sigma|_{\langle id, c\rangle \mathrel{\overline{\ll}} (p\otimes \mathbf{1})})\Big) \\ &= (\pi_2 \circ \langle id, c\rangle) \gg\!= (\sigma|_{(id \mathrel{\overline{\ll}} p)\&(c \mathrel{\overline{\ll}} \mathbf{1})}) \qquad \text{by Exercises 6.1.11 and 4.3.8} \\ &= c \gg\!= (\sigma|_p) \\ &(\langle id, c\rangle \gg\!= \sigma)|_{\mathbf{1}\otimes q}[1,0] \\ &\overset{(6.2)}{=} \pi_1 \gg\!= \Big(\langle id, c\rangle|_{\mathbf{1}\otimes q} \gg\!= (\sigma|_{\langle id, c\rangle \mathrel{\overline{\ll}} (\mathbf{1}\otimes q)})\Big) \\ &= (\pi_1 \circ \langle id, c|_q\rangle) \gg\!= (\sigma|_{(id \mathrel{\overline{\ll}} \mathbf{1})\&(c \mathrel{\overline{\ll}} q)}) \\ &= \sigma|_{c \mathrel{\overline{\ll}} q}. \end{aligned}$$

❑

This result will help us how to perform inference in a Bayesian network, see Section 7.10.

Exercises

6.3.1 We consider some disease with an *a priori* probability (or 'prevalence') of 1%. There is a test for the disease with the following characteristics.

- ('sensitivity') If someone has the disease, then the test is positive with probability of 90%.
- ('specificity') If someone does not have the disease, there is a 95% chance that the test is negative.

1 Take as disease space $D = \{d, d^\perp\}$; describe the prior as a distribution on D;

2 Take as test space $T = \{p, n\}$ and describe the combined sensitivity and specificity as a channel $c\colon D \nrightarrow T$;

3 Show that the predicted positive test probability is almost 6%.

4 Assume that a test comes out positive. Use backward reasoning to prove that the probability of having the disease (the posterior, or 'Positive Predictive Value', PPV) is then a bit more than 15% (to be precise: $\frac{18}{117}$). Explain why it is so low — remembering Example 6.3.3.

6.3.2 In the context of the previous exercise we can derive the familiar formulas for Postive Predictive Value (PPV) and Negative Predictive Value (NPV). Let's assume we have prevalence (prior) given by $\omega = p|d\rangle + (1-p)|d^{\perp}\rangle$ with parameter $p \in [0,1]$ and a channel channel $c\colon \{d, d^{\perp}\} \nrightarrow \{p, n\}$ with parameters:

$$\begin{array}{lll} \text{sensitivity:} & c(d) = se|p\rangle + (1-se)|n\rangle \\ \text{specificity:} & c(d^{\perp}) = (1-sp)|p\rangle + sp|n\rangle. \end{array}$$

Check that:

$$PPV \coloneqq \omega|_{c \nrightarrow \mathbf{1}_p}(d) = \frac{p \cdot se}{p \cdot se + (1-p) \cdot (1-sp)}.$$

This is commonly expressed in medical textbooks as:

$$PPV = \frac{prevalence \cdot sensitivity}{prevalence \cdot sensitivity + (1 - prevalence) \cdot (1 - specificity)}.$$

Check similarly that:

$$NPV \coloneqq \omega|_{c \nrightarrow \mathbf{1}_n}(d^{\perp}) = \frac{(1-p) \cdot sp}{p \cdot (1-se) + (1-p) \cdot sp}.$$

As an aside, the (positive) *likelihood ratio* LR is the fraction:

$$LR \coloneqq \frac{c(d)(p)}{c(d^{\perp})(p)} = \frac{se}{1-sp}.$$

6.3.3 Give a channel-based analysis and answer to the following question from [144, Chap. I, Exc. 39].

> Stores A, B, and C have 50, 75, and 100 employees, and respectively, 50, 60, and 70 percent of these are women. Resignations are equally likely among all employees, regardless of sex. One employee resigns and this is a woman. What is the probability that she works in store C?

6.3.4 The multinomial and hypergeometric charts in Figures 6.3 and 6.4 are very similar. Still, if we look at the probability for the situation when there are 40 fish in the pond, there is clear difference. Give a conceptual explation for this difference.

6.3.5 The following situation about the relationship between eating hamburgers and having Kreuzfeld-Jacob disease is insprired by [8, §1.2]. We have two sets: $E = \{H, H^\perp\}$ about eating Hamburgers (or not), and $D = \{K, K^\perp\}$ about having Kreuzfeld-Jacob disease (or not). The following distributions on these sets are given: half of the people eat hamburgers, and only one in hundred thousand have Kreuzfeld-Jacob disease, which we write as:

$$\omega = \tfrac{1}{2}|H\rangle + \tfrac{1}{2}|H^\perp\rangle \qquad \text{and} \qquad \sigma = \tfrac{1}{100{,}000}|K\rangle + \tfrac{99{,}999}{100{,}000}|K^\perp\rangle.$$

1 Suppose that we know that 90% of the people who have Kreuzfeld-Jacob disease eat Hamburgers. Use this additional information to define a channel $c\colon D \nrightarrow E$ with $c \gg\!= \sigma = \omega$.

2 Compute the probability of getting Kreuzfeld-Jacob for someone eating hamburgers (via backward inference).

6.3.6 Consider in the context of Example 6.3.8 a negative Covid-19 test obtained after 2 days, via the distribution $\varphi_2 = \frac{1}{4}|1\rangle + \frac{1}{2}|2\rangle + \frac{1}{4}|3\rangle$, assuming a uniform disease prior ω. Show that the posterior 'days' distribution is:

$$\begin{aligned}\Big((\omega \otimes \varphi_2)\big|_{t =\!\ll \mathbf{1}_n}\Big)[0,1] &= \tfrac{199}{726}|1\rangle + \tfrac{179}{363}|2\rangle + \tfrac{169}{726}|3\rangle \\ &\approx 0.274|1\rangle + 0.493|2\rangle + 0.233|3\rangle.\end{aligned}$$

Explain why there a shift 'forward', making the earlier days more likely in this posterior — with respect to the prior φ_2.

6.3.7 Consider the following challenge, copied from [153].

(i) I have forgotten what day it is.
(ii) There are ten buses per hour in the week and three buses per hour at the weekend.
(iii) I observe four buses in a given hour.
(iv) What is the probability that it is the weekend?

Let $W = \{wd, we\}$ be a set with elements wd for weekday and we for weekend, with prior distribution $\frac{5}{7}|wd\rangle + \frac{2}{7}|we\rangle$. Use the Poisson distribution to define a channel $bus\colon W \to \mathcal{D}_\infty(\mathbb{N})$ and use it to answer the above question via backward inference.

6.3.8 Let $c\colon X \nrightarrow Y$ be a channel with state $\omega \in \mathcal{D}(X)$ and factors $p \in Fact(X)$, $q \in Fact(Y)$. Check that:

$$c|_q \gg\!= \Big(\omega\big|_{p \,\&\, (c \prec\!\ll q)}\Big) = \Big(c \gg\!= (\omega|_p)\Big)\big|_q.$$

Describe what this means for $p = \mathbf{1}$.

6.3.9 Let $p \in Fact(X)$ and $q \in Fact(Y)$ be two factors on spaces X, Y.

1 Show that for two channels $c \colon Z \nrightarrow X$ and $d \colon Z \nrightarrow Y$ with a joint state on their (common) domain $\sigma \in \mathcal{D}(Z)$ one has:

$$\begin{aligned}(\langle c, d\rangle \gg\!= \sigma)\big|_{p\otimes q}[1,0] &= (c \gg\!= (\sigma|_{d \prec\!\ll q}))\big|_p \\ (\langle c, d\rangle \gg\!= \sigma)\big|_{p\otimes q}[0,1] &= (d \gg\!= (\sigma|_{c \prec\!\ll p}))\big|_q\end{aligned}$$

2 For channels $e \colon U \nrightarrow X$ and $d \colon V \nrightarrow Y$ with a joint state $\omega \in \mathcal{D}(U \times V)$ one has:

$$\begin{aligned}((e \otimes f) \gg\!= \omega)\big|_{p\otimes q}[1,0] &= \Big(e \gg\!= (\omega\big|_{\mathbf{1}\otimes(f \prec\!\ll q)}[1,0])\Big)\big|_p \\ ((e \otimes f) \gg\!= \omega)\big|_{p\otimes q}[0,1] &= \Big(f \gg\!= (\omega\big|_{(e \prec\!\ll p)\otimes\mathbf{1}}[0,1])\Big)\big|_q.\end{aligned}$$

6.4 Discretisation, and coin bias learning

So far we have been using *discrete* probability distributions, with a finite domain. We shall be looking at *continuous* distributions later on, in Chapter **??**. In the meantime we can approximate continuous distributions by discretisation, namely by chopping them up into finitely many parts, like in Riemann integration. In fact, computers doing computation in continuous probability perform such fine-grained discretisation. This section first introduces such discretisation and then uses it in an extensive example on learning the bias of a coin via successive backward inference. It is classical illustration.

We start with discretisation.

Definition 6.4.1. Let $a, b \in \mathbb{R}$ with $a < b$ and $N \in \mathbb{N}$ with $N > 0$ be given.

1 We write $[a, b]_N \subseteq [a, b] \subseteq \mathbb{R}$ for the interval $[a, b]$ reduced to N elements:

$$\begin{aligned}[a, b]_N &:= \{a + (i + \tfrac{1}{2})s \mid i \in N\} \qquad \text{where} \quad s := \tfrac{b-a}{N} \\ &= \{a + \tfrac{1}{2}s,\ a + \tfrac{3}{2}s, \ldots, a + \tfrac{2N-1}{2}s\} \\ &= \{a + \tfrac{1}{2}s,\ a + \tfrac{3}{2}s, \ldots, b - \tfrac{1}{2}s\}.\end{aligned}$$

2 Let $f \colon S \to \mathbb{R}_{\geq 0}$ be a function, defined on a finite subset $S \subseteq \mathbb{R}$. We write $\mathit{Disc}(f, S) \in \mathcal{D}(S)$ for the discrete distribution defined as:

$$\mathit{Disc}(f, S) := \sum_{x \in S} \tfrac{f(x)}{t}\big|x\big\rangle \qquad \text{where} \quad t := \textstyle\sum_{x\in S} f(x).$$

Often we combine the notations from these two items and use discretised states of the form $\mathit{Disc}(f, [a, b]_N)$.

To see an example of item (1), consider the interval $[1, 2]$ with $N = 3$. The step size s is then $s = \frac{2-1}{3} = \frac{1}{3}$, so that:

$$[1, 2]_3 = \{ 1 + \tfrac{1}{2} \cdot \tfrac{1}{3},\ 1 + \tfrac{3}{2} \cdot \tfrac{1}{3},\ 1 + \tfrac{5}{2} \cdot \tfrac{1}{3} \} = \{ 1 + \tfrac{1}{6},\ 1 + \tfrac{1}{2},\ 1 + \tfrac{5}{6} \}$$

We choose to use internal points only and exclude the end-points in this finite subset since the end-points sometimes give rise to boundary problems, with functions being undefined. When N goes to infinity, the smallest and largest elements in $[a, b]_N$ will approximate the end-points a and b — from above and from below, respectively.

The 'total' number t in item (2) normalises the formal sum and ensures that the multiplicities add up to one. In this way we can define a uniform distribution on $[a, b]_N$ as $\mathit{unif}_{[a,b]_N}$, like before, or alternatively as $\mathit{Disc}(\mathbf{1}, [a, b]_N)$, where $\mathbf{1}$ is the constant-one function.

Example 6.4.2. We look at the following classical question: suppose we are given a coin with an unknown bias, we flip it eight times, and observe the following list of heads (H) and tails (T):

$$[T, H, H, H, T, T, H, H].$$

What can we then say about the bias of the coin?

The frequentist approach that we have seen in Section 2.3 would turn the above list into a multiset and then into a distribution, by frequentist learning, see also Diagram (2.16). This gives:

$$[T, H, H, H, T, T, H, H] \longmapsto 5|H\rangle + 3|T\rangle \longmapsto \tfrac{5}{8}|H\rangle + \tfrac{3}{8}|T\rangle.$$

Here we do not use this frequentist approach to learning the bias parameter, but take a Bayesian route. We assume that the bias parameter itself is given by a distribution, describing the likelihoods of various bias values. We assume no prior knowledge and therefore start from the uniform distribution. It will be updated based on successive observations, using the technique of backward inference, see Definition 6.3.1 (2).

The bias b of a coin is a number in the unit interval $[0, 1]$, giving rise to a coin distribution $\mathit{flip}(b) = b|H\rangle + (1 - b)|T\rangle$. Thus we can see flip as a channel $\mathit{flip} \colon [0, 1] \nrightarrow \{H, T\}$. At this stage we avoid continuous distributions and discretise the unit interval. We choose $N = 100$ in the chop up, giving as underlying space $[0, 1]_N$ with N points, on which we take the uniform distribution unif as prior:

$$\begin{aligned} \mathit{unif} \coloneqq \mathit{Disc}(\mathbf{1}, [0, 1]_N) &= \sum_{x \in [0,1]_N} \tfrac{1}{N}|x\rangle = \sum_{0 \leq i < N} \tfrac{1}{N}\big|\tfrac{2i+1}{2N}\big\rangle \\ &= \tfrac{1}{N}\big|\tfrac{1}{2N}\big\rangle + \tfrac{1}{N}\big|\tfrac{3}{2N}\big\rangle + \cdots + \tfrac{1}{N}\big|\tfrac{2N-1}{2N}\big\rangle. \end{aligned}$$

We use the *flip* operation as a channel, restricted to the discretised space:

$$[0,1]_N \xrightarrow{\;\mathit{flip}\;} \{H,T\} \qquad \text{given by} \qquad \mathit{flip}(b) \;=\; b|H\rangle + (1-b)|T\rangle.$$

There are the two (sharp, point) predicates $\mathbf{1}_H$ and $\mathbf{1}_T$ on the codomain $\{H,T\}$. It is not hard to show, see Exercise 6.4.2 below, that:

$$\mathit{flip} \gg\!= \mathit{unif} \models \mathbf{1}_H \;=\; \mathit{flip} \gg\!= \mathit{unif} \models \mathbf{1}_T \;=\; \tfrac{1}{2}.$$

Predicate transformation along *flip* yields two predicates on $[0,1]_N$ given by:

$$(\mathit{flip} =\!\ll \mathbf{1}_H)(r) \;=\; r \qquad \text{and} \qquad (\mathit{flip} =\!\ll \mathbf{1}_T)(r) \;=\; 1 - r.$$

Given the above sequence of head/tail observations $[T,H,H,H,T,T,H,H]$, we perform successive predicate transformations $\mathit{flip} =\!\ll \mathbf{1}_{(-)}$ and update the prior (uniform) state accordingly. This gives, via Lemma 6.1.6 (3),

$$\begin{aligned}
&\mathit{unif}|_{\mathit{flip} =\!\ll \mathbf{1}_T} \\
&\mathit{unif}|_{\mathit{flip} =\!\ll \mathbf{1}_T}|_{\mathit{flip} =\!\ll \mathbf{1}_H} \;=\; \mathit{unif}|_{(\mathit{flip} =\!\ll \mathbf{1}_T)\&(\mathit{flip} =\!\ll \mathbf{1}_H)} \;=\; \mathit{unif}|_{(\mathit{flip} =\!\ll \mathbf{1}_H)\&(\mathit{flip} =\!\ll \mathbf{1}_T)} \\
&\mathit{unif}|_{\mathit{flip} =\!\ll \mathbf{1}_T}|_{\mathit{flip} =\!\ll \mathbf{1}_H}|_{\mathit{flip} =\!\ll \mathbf{1}_H} \;=\; \mathit{unif}|_{(\mathit{flip} =\!\ll \mathbf{1}_T)\&(\mathit{flip} =\!\ll \mathbf{1}_H)\&(\mathit{flip} =\!\ll \mathbf{1}_H)} \\
&\qquad\qquad\qquad\qquad\qquad\quad = \mathit{unif}|_{(\mathit{flip} =\!\ll \mathbf{1}_H)^2\&(\mathit{flip} =\!\ll \mathbf{1}_T)} \\
&\qquad\vdots \\
&\mathit{unif}|_{(\mathit{flip} =\!\ll \mathbf{1}_H)^5\&(\mathit{flip} =\!\ll \mathbf{1}_T)^3}
\end{aligned}$$

An overview of the resulting distributions is given in Figure 6.5. These distributions approximate (continuous) beta distributions, see Section **??** later on. These beta functions form a smoothed out version of the bar charts in Figure 6.5. We shall look more systematically into 'learning along a channel' in Section **??**, where the current coin-bias computation re-appears in Example **??**.

After these eight updates, let's write $\rho = \mathit{unif}|_{(\mathit{flip} =\!\ll \mathbf{1}_H)^5\&(\mathit{flip} =\!\ll \mathbf{1}_T)^3}$ for the resulting distribution. We now ask three questions.

1 Where does ρ reach its highest value, and what is it? The answers are given by:

$$\operatorname{argmax}(\rho) \;=\; \{\tfrac{25}{40}\} \qquad \text{with} \qquad \rho(\tfrac{25}{40}) \;\approx\; 0.025347.$$

2 What is the predicted coin distribution? The outcome, with truncated multiplicities, is:

$$\mathit{flip} \gg\!= \rho \;=\; 0.6|1\rangle + 0.4|0\rangle.$$

3 What is the expected value of ρ? It is:

$$\mathit{mean}(\rho) \;=\; 0.6 \;=\; \rho \models \mathit{flip} =\!\ll \mathbf{1}_H.$$

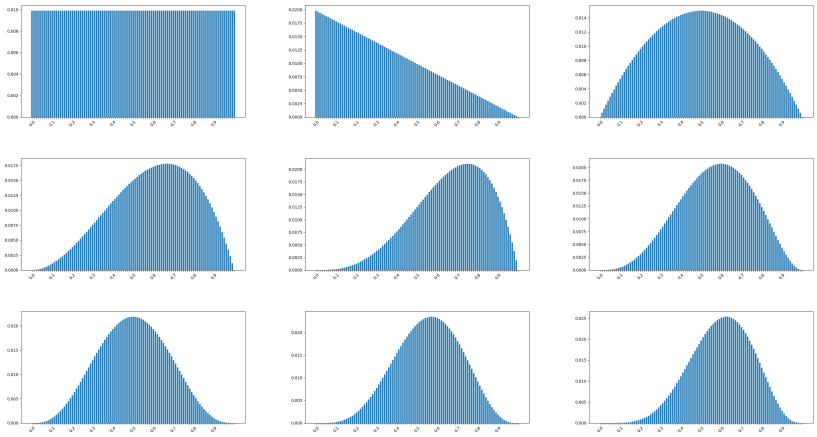


Figure 6.5 Coin bias distributions arising from the prior uniform distribution by successive updates after coin observations $[0, 1, 1, 1, 0, 0, 1, 1]$.

For mathematical reasons[1] the exact outcome is 0.6. However, we have used approximation via discretisation. The value computed with this discretisation is 0.599999985316273. We can conclude that chopping the unit interval up with $N = 100$ already gives a fairly good approximation.

We conclude with a 'conjugate prior' property in this situation. It is a discrete instance of a phenomenon that will be described more systematically later on, see Section **??**, in continuous probability.

We define, parameterised by $N \in \mathbb{N}$, the channel $\beta_N \colon \mathbb{N} \times \mathbb{N} \nrightarrow [0,1]_N$ as normalisation:

$$\begin{aligned} \beta_N(a,b) &:= \mathit{Flrn}\left(\sum_{r \in [0,1]_N} r^a \cdot (1-r)^b \big| r \right) \\ &= \sum_{r \in [0,1]_N} \frac{r^a \cdot (1-r)^b}{\sum_{s \in [0,1]_N} s^a \cdot (1-s)^b} \big| r \big\rangle. \end{aligned} \tag{6.3}$$

Then we have the following results.

Proposition 6.4.3. *Let $N \in \mathbb{N}$ be the the discretisation parameter, used in the chopped subspace $[0,1]_N \subseteq [0,1]$.*

[1] The distribution ρ is an approximation of the probability density function $\beta(5,3)$, which has mean $\frac{5+1}{(5+1)+(3+1)} = 0.6$.

1 For $a, b \in \mathbb{N}$ the above distribution $\beta_N(a, b)$ satisfies:

$$\beta_N(a, b) \;=\; \mathit{unif}\big|_{(\mathit{flip} \nsim \mathbf{1}_H)^a \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^b}.$$

2 For additional numbers $n, m \in \mathbb{N}$ one has:

$$\beta_N(a, b)\big|_{(\mathit{flip} \nsim \mathbf{1}_H)^n \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^m} \;=\; \beta_N(a + n, b + m).$$

This last equation says that the β_N distributions are closed under updating via *flip*. This is the essence of the fact that the β_N channel is 'conjugate prior' to the *flip* channel, see Section **??** for more details. For now we note that this relationship is convenient because it means that we don't have to perform all the state updates explicitly; instead we can just adapt the inputs a, b of the channel β_N. These inputs are often called hyperparameters.

Proof. 1 For $r \in D$ we have:

$$\begin{aligned}
\mathit{unif}\big|_{(\mathit{flip} \nsim \mathbf{1}_H)^a \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^b}(r) \;&=\; \frac{\mathit{unif}(r) \cdot (\mathit{flip} \nsim \mathbf{1}_H)^a(r) \cdot (\mathit{flip} \nsim \mathbf{1}_T)^b(r)}{\mathit{unif} \models (\mathit{flip} \nsim \mathbf{1}_H)^a \;\&\; (\mathit{flip} \nsim \mathbf{1}_T)^b} \\
&= \frac{1/N \cdot r^a \cdot (1 - r)^b}{\sum_{s \in D} 1/N \cdot s^a \cdot (1 - s)^b} \\
&= \beta_N(a, b).
\end{aligned}$$

2 We use this result and Lemma 6.1.6 (3) in:

$$\begin{aligned}
&\beta_N(a, b)\big|_{(\mathit{flip} \nsim \mathbf{1}_H)^n \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^m} \\
&= \mathit{unif}\big|_{(\mathit{flip} \nsim \mathbf{1}_H)^a \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^b}\big|_{(\mathit{flip} \nsim \mathbf{1}_H)^n \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^m} \\
&= \mathit{unif}\big|_{(\mathit{flip} \nsim \mathbf{1}_H)^a \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^b \,\&\, (\mathit{flip} \nsim \mathbf{1}_H)^n \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^m} \\
&= \mathit{unif}\big|_{(\mathit{flip} \nsim \mathbf{1}_H)^{a+n} \,\&\, (\mathit{flip} \nsim \mathbf{1}_T)^{b+m}} \\
&= \beta_N(a + n, b + m).
\end{aligned}$$

❑

6.4.1 Discretisation of states

In the beginning of this section we have described how to chop up intervals $[a, b]$ of real numbers into a discrete sample space. We have used this in particular for the unit interval $[0, 1]$ which we used as space for a coin bias. Since there is an isomorphism $[0, 1] \cong \mathcal{D}(2)$, this discretisation of $[0, 1]$ might as well be seen as a discretisation of the set of states on $2 = \{0, 1\}$. Can we do such discretisation more generally, for sets of states $\mathcal{D}(X)$? There is an easy way to do so via normalisation of natural multisets.

Recall that we write $\mathcal{N}[K](X)$ for the set of natural multisets — with natural

numbers as multiplicities — of size K. We recall from Theorem 4.5.7 (2) that we write:

$$\mathcal{D}[K](X) \;=\; \{\mathit{Flrn}(\varphi) \mid \varphi \in \mathcal{N}[K]\} \;=\; \{\tfrac{1}{N} \cdot \varphi \mid \varphi \in \mathcal{N}[K]\}.$$

Recall also — from Proposition 1.7.4 — that if the set X has n elements, then $\mathcal{N}[K](X)$ contains $\left(\!\binom{n}{K}\!\right)$ multisets, so that $\mathcal{D}[K](X)$ contains $\left(\!\binom{n}{K}\!\right)$ distributions. For instance, when $X = \{a, b\}$, then $\mathcal{M}[5](X)$ contains the multisets:

$$5|a\rangle,\ 4|a\rangle + 1|b\rangle,\ 3|a\rangle + 2|b\rangle,\ 2|a\rangle + 3|b\rangle,\ 1|a\rangle + 4|b\rangle,\ 5|b\rangle.$$

Hence, $\mathcal{D}[5](\{a, b\})$ contains the distributions:

$$1|a\rangle,\ \tfrac{4}{5}|a\rangle + \tfrac{1}{5}|b\rangle,\ \tfrac{3}{5}|a\rangle + \tfrac{2}{5}|b\rangle,\ \tfrac{2}{5}|a\rangle + \tfrac{3}{5}|b\rangle,\ \tfrac{1}{5}|a\rangle + \tfrac{4}{5}|b\rangle,\ 1|b\rangle.$$

By taking large K in $\mathcal{D}[K](X)$ we can obtain fairly good approximations, since the union of these subsets $\mathcal{D}[K](X) \subseteq \mathcal{D}(X)$ is dense, by Theorem 4.5.7 (2).

Example 6.4.4. Consider an election with three candidates a, b, c, which we collect in a set X of candidates $X = \{a, b, c\}$. A candidate wins if (s)he gets more than 50% of the votes. A poll has been conducted among 100 possible voters, giving the following preferences.

candidates	a	b	c
poll numbers	52	28	20

What is the probability that candidate a then wins in the election? Notice that these are poll numbers, before the election, not actual votes. If they were actual votes, candidate a gets more than 50%, and thus wins. However, these are poll numbers, from which we like to learn about voter preferences.

A state in $\mathcal{D}(X)$ is used as distribution of preferences over the three candidates in $X = \{a, b, c\}$. In this example we discretise the set of states and work with the finite subset $\mathcal{D}[K](X) \hookrightarrow \mathcal{D}(X)$, for a number K. There is no initial knowledge about voter preferences, so our prior is the uniform distribution over these preference distributions:

$$\upsilon_K \;\coloneqq\; \mathit{unif}_{\mathcal{D}[K](X)} \in \mathcal{D}\Big(\mathcal{D}[K](X)\Big).$$

We use the following sharp predicates on $\mathcal{D}[K](X)$, called: aw for a wins, bw for b wins, cw for c wins, and nw for no-one wins. The are defined on

$\sigma \in \mathcal{D}[K](X)$ as:

$$\begin{aligned} aw(\sigma) = 1 &\iff \sigma(a) > \tfrac{1}{2} \\ bw(\sigma) = 1 &\iff \sigma(b) > \tfrac{1}{2} \\ cw(\sigma) = 1 &\iff \sigma(c) > \tfrac{1}{2} \\ aw(\sigma) = 1 &\iff \sigma(a) \leq \tfrac{1}{2} \text{ and } \sigma(b) \leq \tfrac{1}{2} \text{ and } \sigma(c) \leq \tfrac{1}{2}. \end{aligned}$$

The prior validities:

$$\upsilon_K \models aw \qquad \upsilon_K \models bw \qquad \upsilon_K \models cw \qquad \upsilon_K \models nw$$

all approximate $\frac{1}{4}$, as K goes to infinity.

We use the inclusion $\mathcal{D}[K](X) \hookrightarrow \mathcal{D}(X)$ as a channel $i \colon \mathcal{D}[K](X) \nrightarrow X$. Like in the above coin scenario, we perform successive backward inferences, using the above poll numbers, to obtain a posterior ρ_K obtained via updates:

$$\rho_K \coloneqq \upsilon_K|_{(i =\!\ll \mathbf{1}_a)^{52}}|_{(i =\!\ll \mathbf{1}_b)^{28}}|_{(i =\!\ll \mathbf{1}_c)^{20}} = \upsilon_K|_{(i =\!\ll \mathbf{1}_a)^{52} \,\&\, (i =\!\ll \mathbf{1}_b)^{28} \,\&\, (i =\!\ll \mathbf{1}_c)^{20}}.$$

We illustrate that the posterior probability $\rho_K \models aw$ takes the following values, for several numbers K.

	$N = 100$	$N = 500$	$N = 1000$
probability that a wins	0.578	0.609	0.613

Clearly, these probabilities are approximations. When modeled via continuous probability theory, see Chapter **??**, the probability that a wins can be calculated more accurately as 0.617. This illustrates that the above discretisations of states work reasonably well.

To give a bit more perspective, the posterior probability that b or c wins with the above poll numbers is close to zero. The probability that no-one wins is substantial, namely almost 0.39.

Exercises

6.4.1 Recall the the N-element set $[a, b]_N$ from Definition 6.4.1 (1).

1 Show that its largest element is $b - \frac{1}{2}s$, where $s = \frac{b-a}{N}$.

2 Prove that $\sum_{x \in [a,b]_N} x = \frac{N(a+b)}{2}$. Remember: $\sum_{i \in N} i = \frac{N(N-1)}{2}$.

6.4.2 Consider coin parameter learning in Example 6.4.2.

1 Show that the predicition in the prior (uniform) state *unif* on $[0, 1]_N$ gives a fair coin, *i.e.*

$$\mathit{flip} \gg\!= \mathit{unif} = \tfrac{1}{2}|1\rangle + \tfrac{1}{2}|0\rangle.$$

This equality is independent of N.

2 Prove that, also independently of N,

$$\mathit{unif} \models \mathit{flip} \mathrel{-\!\!\ll} \mathbf{1}_H \;=\; \tfrac{1}{2}.$$

3 Show next that:

$$\mathit{unif}|_{\mathit{flip} \mathrel{-\!\!\ll} \mathbf{1}_H} \;=\; \sum\nolimits_{x \in [0,1]_N} \tfrac{2x}{N} |\, x \,\rangle.$$

4 Use the 'square pyramidal' formula $\sum_{i \in N} i^2 = \frac{N(N-1)(2N-1)}{6}$ to prove that:

$$(\mathit{flip} \gg\!= (\mathit{unif}|_{\mathit{flip} \mathrel{-\!\!\ll} \mathbf{1}_H}))(1) \;=\; \tfrac{2}{3} - \tfrac{1}{6N^2}.$$

Conclude that $\mathit{flip} \gg\!= (\mathit{unif}|_{\mathit{flip} \mathrel{-\!\!\ll} \mathbf{1}_H})$ approaches $\frac{2}{3}|1\rangle + \frac{1}{3}|0\rangle$ as N goes to infinity.

6.4.3 Check that the probabilities $(\mathit{flip} \gg\!= \rho)(1)$ and $\mathit{mean}(\rho)$ are the same, in Example 6.4.2.

6.4.4 Check that $\beta_N(0, 0)$ in (6.3) is the uniform distribution *unif* on $[0, 1]_N$.

6.4.5 Consider the discretised beta channel $\beta_N \colon \mathbb{N} \times \mathbb{N} \to \mathcal{D}(X)$ in the following diagram:

$$\begin{array}{ccc} (\mathbb{N} \times \mathbb{N}) \times \mathcal{N}(\{H, T\}) & \xrightarrow{\beta_N \times id} & \mathcal{D}([0,1]_N) \times \mathcal{N}(\{H, T\}) \\ {\scriptstyle add}\downarrow & & \downarrow{\scriptstyle infer} \\ \mathbb{N} \times \mathbb{N} & \xrightarrow{\qquad\qquad \beta_N \qquad\qquad} & \mathcal{D}([0,1]_N) \end{array}$$

where the maps *add* and *infer* are given by:

$$\begin{aligned} \mathit{add}(a,\, b,\, n|H\rangle + m|T\rangle) &= (a + n, b + m) \\ \mathit{infer}(\omega,\, n|H\rangle + m|T\rangle) &= \omega|_{(\mathit{flip} \mathrel{-\!\!\ll} \mathbf{1}_H)^n \,\&\, (\mathit{flip} \mathrel{-\!\!\ll} \mathbf{1}_T)^m}. \end{aligned}$$

Show that:

1 The map *add* is an action of the monoid $\mathcal{N}(\{H, T\})$ on $\mathbb{N} \times \mathbb{N}$.

2 Show that the above rectangle commutes — and thus that β_N is a homomorphism of monoid actions. (This description of a conjugate prior relationship as monoid action comes from [81].)

6.4.6 We accept, without proof, the following equation. For $a, b \in \mathbb{N}$,

$$\lim_{N \to \infty} \frac{1}{N} \cdot \sum_{r \in [0,1]_N} r^a \cdot (1 - r)^b \;=\; \frac{a! \cdot b!}{(a + b + 1)!}. \tag{$*$}$$

Use $(*)$ to prove that the binary Pólya distribution can be obtained from the binomial, via the limit of the discretised β distribution (6.3):

$$\begin{aligned} &\lim_{N \to \infty} \Big(\mathit{bn}[K] \gg\!= \beta_N(a, b) \Big)(i) \\ &= \mathit{pl}[K]\Big((a+1)|H\rangle + (b+1)|T\rangle\Big)\Big(i|H\rangle + (K-i)|T\rangle\Big). \end{aligned}$$

Later, in **??** we shall see a proper continuous version of this result.

6.4.7 Redo the coin bias learning from the beginning of this section in the style of Example 6.4.4, via discretisation of states in $\mathcal{D}(2)$.

6.5 Inference in Bayesian networks

In previous sections we have seen several examples of channel-based inference, in forward and backward form. This section shows how to apply these inference methods to Bayesian networks, via an example that is often used in the literature: the 'Asia' Bayesian network, originally from [110]. It captures the situation of patients with a certain probability of smoking and of an earlier visit to Asia; this influences certain lung diseases and the outcome of an xray test. Later on, in Section 7.10, we shall look more systematically at inference in Bayesian networks.

The Bayesian network example considered here is described in several steps: Figure 6.6 gives the underlying graph structure, in the style of Figure 2.4, introduced at the end of Section 2.6, with diagonals and types of wires written explicitly. Figure 6.7 gives the conditional probability tables associated with the nodes of this network. The way in which these tables are written is different from Section 2.6: we now only write the probabilities $r \in [0, 1]$ and omit the values $1 - r$. Next, these conditional probability tables are described in Figure 6.8 as states $\mathit{smoke} \in \mathcal{D}(S)$ and $\mathit{asia} \in \mathcal{D}(A)$ and as channels $\mathit{lung} \colon S \nrightarrow L$, $\mathit{tub} \colon A \nrightarrow T$, $\mathit{bronc} \colon S \nrightarrow B$, $\mathit{xray} \colon E \nrightarrow X$, $\mathit{dysp} \colon B \times E \nrightarrow D$, $\mathit{either} \colon L \times T \nrightarrow E$.

Our aim in this section is illustrate channel-based inference in this Bayesian network. It is not so much the actual outcomes that we are interested in, but more the systematic methodology that is used to obtain these outcomes.

Probability of lung cancer, given no bronchitis

Let's start with the question: what is the probability that someone has lung cancer, given that this person does not have bronchitis. The latter information is the evidence. It takes the form of a singleton predicate $\mathbf{1}_{b^\perp} = (\mathbf{1}_b)^\perp \colon B \to [0, 1]$ on the set $B = \{b, b^\perp\}$ used for presence and absence of bronchitis.

In order to obtain this updated probability of lung cancer we 'follow the graph'. In Figure 6.6 we see that we can transform (pull back) the evidence along the bronchitis channel $\mathit{bronc} \colon A \nrightarrow B$, and obtain a predicate $\mathit{bronc} \mathrel{=\!\!\ll} \mathbf{1}_{b^\perp}$ on A. The latter can be used to update the asia distribution on A. Subsequently, we can push the updated distribution forward along the lung channel

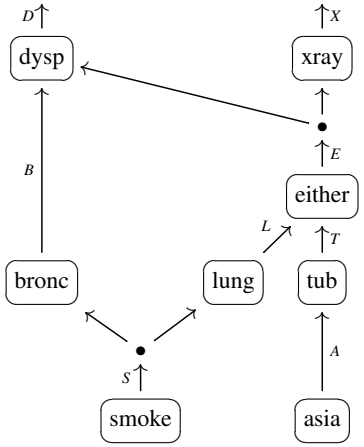


Figure 6.6 The graph of the Asia Bayesian network, with node abbreviations: bronc = bronchitis, dysp = dyspnea, lung = lung cancer, tub = tuberculosis. The wires all have 2-element sets of the form $A = \{a, a^\perp\}$ as types.

lung: $A \nrightarrow L$ via state transformation. Thus we follow the 'V' shape in the relevant part of the graph.

Combining this down-update-up steps gives the required outcome:

$$\mathit{lung} \gg\!= (\mathit{smoke}\big|_{\mathit{bronc} \,\overline{\ll}\, \mathbf{1}_{b^\perp}}) \;=\; 0.0427|\ell\rangle + 0.9573|\ell^\perp\rangle.$$

We see that this calculation combines forward and backward inference, see Definition 6.3.1.

Probability of smoking, given a positive xray

In Figures 6.7 and 6.8 we see a prior smoking probability of 50%. We like to know what this probability becomes if we have evidence of a positive xray. The latter is given by the point predicate $\mathbf{1}_x \in \mathit{Pred}(X)$ for $X = \{x, x^\perp\}$.

Now there is a long path from xray to smoking, see Figure 6.6 that we need to use for (backward) predicate transformation. Along the way there is a complication, namely that the node 'either' has two parent nodes, so that pulling back along the *either* channel yields a predicate on the product set $L \times T$. The only sensible thing to do is continue predicate transformation down downwards, but now with the parallel product channel $\mathit{lung} \otimes \mathit{tub} : S \times A \nrightarrow L \times T$. The resulting predicate on $S \times A$ can be used to update the product state $\mathit{smoke} \otimes \mathit{asia}$. Then we can take the first marginal to obtain the desired outcome.

P(smoke)
0.5

smoke	P(lung)
s	0.1
$s^\perp$	0.01

smoke	P(bronc)
s	0.6
$s^\perp$	0.3

bronc	either	P(dysp)
b	e	0.9
b	$e^\perp$	0.7
$b^\perp$	e	0.8
$b^\perp$	$e^\perp$	0.1

P(asia)
0.01

asia	P(tub)
a	0.05
$a^\perp$	0.01

either	P(xray)
e	0.98
$e^\perp$	0.05

lung	tub	P(either)
ℓ	t	1
ℓ	$t^\perp$	1
$\ell^\perp$	t	1
$\ell^\perp$	$t^\perp$	0

Figure 6.7 The conditional probability tables of the Asia Bayesian network

$$
\begin{aligned}
\mathit{smoke} &= 0.5|s\rangle + 0.5|s^\perp\rangle \\
\mathit{lung}(s) &= 0.1|\ell\rangle + 0.5|\ell^\perp\rangle \\
\mathit{lung}(s^\perp) &= 0.3|\ell\rangle + 0.7|\ell^\perp\rangle \\
\mathit{bronc}(s) &= 0.6|b\rangle + 0.4|b^\perp\rangle \\
\mathit{bronc}(s^\perp) &= 0.3|b\rangle + 0.7|b^\perp\rangle \\
\mathit{dysp}(b,e) &= 0.9|d\rangle + 0.1|d^\perp\rangle \\
\mathit{dysp}(b,e^\perp) &= 0.7|d\rangle + 0.3|d^\perp\rangle \\
\mathit{dysp}(b^\perp,e) &= 0.8|d\rangle + 0.2|d^\perp\rangle \\
\mathit{dysp}(b^\perp,e^\perp) &= 0.1|d\rangle + 0.9|d^\perp\rangle
\end{aligned}
\qquad
\begin{aligned}
\mathit{asia} &= 0.01|a\rangle + 0.99|a^\perp\rangle \\
\mathit{tub}(a) &= 0.05|t\rangle + 0.95|t^\perp\rangle \\
\mathit{tub}(a^\perp) &= 0.01|t\rangle + 0.99|t^\perp\rangle \\
\mathit{xray}(e) &= 0.98|x\rangle + 0.02|x^\perp\rangle \\
\mathit{xray}(e^\perp) &= 0.05|x\rangle + 0.95|x^\perp\rangle \\
\mathit{either}(\ell,t) &= 1|e\rangle \\
\mathit{either}(\ell,t^\perp) &= 1|e\rangle \\
\mathit{either}(\ell^\perp,t) &= 1|e\rangle \\
\mathit{either}(\ell^\perp,t^\perp) &= 1|e^\perp\rangle
\end{aligned}
$$

Figure 6.8 The conditional probability tables from Figure 6.7 described as states and channels.

Thus we compute:

$$
\begin{aligned}
&\Big((\mathit{smoke}\otimes\mathit{asia})\big|_{(\mathit{lung}\otimes\mathit{tub})\,\prec\!\!<\,(\mathit{either}\,\prec\!\!<\,(\mathit{xray}\,\prec\!\!<\,\mathbf{1}_x))}\Big)[1,0] \\
&= \Big((\mathit{smoke}\otimes\mathit{asia})\big|_{(\mathit{xray}\,\circ\,\mathit{either}\,\circ\,(\mathit{lung}\otimes\mathit{tub}))\,\prec\!\!<\,\mathbf{1}_x}\Big)[1,0] \\
&= 0.6878|s\rangle + 0.3122|s^\perp\rangle.
\end{aligned}
$$

Thus, a positive xray makes it more likely — w.r.t. the uniform prior — that the patient smokes — as is to be expected. This is obtained by backward inference.

Probability of lung cancer, given both dyspnoea and tuberculosis

Our next inference challenge involves two evidence predicates, namely $\mathbf{1}_d$ on D for dyspnoea and $\mathbf{1}_t$ on T for tuberculosis. We would like to know the updated lung cancer probability.

The situation looks complicated, because of the 'closed loop' in Figure 6.6. But we can proceed in a straightforward manner and combine evidence via conjunction & at a suitable meeting point. We now more clearly separate the forward and backward stages of the inference process. We first move the prior states forward to a point that includes the set L — the one that we need to marginalise on to get our conclusion. We abbreviate this state on $B \times L \times T$ as:

$$\sigma \coloneqq (\mathit{bronc} \otimes \mathit{lung} \otimes \mathit{id}) \gg\!= ((\Delta \otimes \mathit{tub}) \gg\!= (\mathit{smoke} \otimes \mathit{asia})).$$

Recall that we write Δ for the copy channel, in this expression of type $S \multimap S \times S$.

Going in the backward direction we can form a predicate, called p below, on the set $B \times L \times T$, by predicate transformation *and* conjunction:

$$p \coloneqq (\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}_t) \mathbin{\&} ((\mathit{id} \otimes \mathit{either}) =\!\ll (\mathit{dys} =\!\ll \mathbf{1}_d)).$$

The result that we are after is now obtained via updating and marginalisation:

$$\sigma|_p[0, 1, 0] \;=\; 0.0558|\ell\rangle + 0.9442|\ell^\perp\rangle. \tag{6.4}$$

There is an alternative way to describe the same outcome, using that certain channels can be 'shifted'. In particular, in the definition of the above state σ, the channel *bronc* is used for state transformation. It can also be used in a different role, namely for predicate transformation. We then use a slightly different state, now on $S \times L \times T$,

$$\tau \coloneqq (\mathit{id} \otimes \mathit{lung} \otimes \mathit{id}) \gg\!= ((\Delta \otimes \mathit{tub}) \gg\!= (\mathit{smoke} \otimes \mathit{asia})).$$

The *bronc* channel is now used for predicate transformation in the predicate:

$$q \coloneqq (\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}_t) \mathbin{\&} ((\mathit{bronc} \otimes \mathit{either}) =\!\ll (\mathit{dys} =\!\ll \mathbf{1}_d)).$$

The same updated lung cancer distribution is now obtained as:

$$\tau|_q[0, 1, 0] \;=\; 0.0558|\ell\rangle + 0.9442|\ell^\perp\rangle. \tag{6.5}$$

The reason why the outcomes (6.4) and (6.5) are the same is the topic of Exercise 6.5.2.

We conclude that inference in Bayesian networks can be done compositionally via a combination of forward and backward inference, basically by following the network structure.

Exercises

6.5.1 Consider the wetness Bayesian network from Section 2.6. Write down the channel-based inference formulas for the following inference questions and check the outcomes that are given below.

1 The updated sprinkler distribution, given evidence of a slippery road, is $\frac{63}{260}|b\rangle + \frac{197}{260}|b\rangle$.
2 The updated wet grass distribution, given evidence of a slippery road, is $\frac{4349}{5200}|d\rangle + \frac{851}{5200}|d^{\perp}\rangle$.

(Answers will appear later, in Example 7.3.3, in two different ways.)

6.5.2 Check that the equality of the outcomes in (6.4) and in (6.5) can be explained via Exercises 6.1.7 (2) and 4.3.9.

6.6 Bayesian inversion: the dagger of a channel

So far in this chapter we have seen the power of probabilistic inference, especially in backward form, especially in the many examples in Section 6.3. Applying backward inference to point predicates, as evidence, leads to channel reversal. This is a fundamental construction, known as Bayesian inversion, which we shall describe with a 'dagger' $c^{\dagger} \colon Y \nrightarrow X$, for a channel $c \colon X \nrightarrow Y$. The superscript-dagger notation $(-)^{\dagger}$ is used since Bayesian inversion is similar to the adjoint-transpose of a linear operator A between Hilbert spaces, see [24] (and Exercise 6.6.3 below). This transpose is typically written as $A^{\dagger}$, or also as A^{*}. The dagger notation is more common in quantum theory, and has been formalised in terms of dagger categories [25]. This categorical approach is sketched in Section 7.8 below.

There is much to say about Bayesian inversion and dagger channels. In the remainder of this chapter we introduce the definition, together with some basic results, and apply inversion in probabilistic reasoning, using a new update rule due to Jeffrey. In Chapter 7 we will re-introduce Bayesian inversion as a special case of disintegration, in the context of a graphical calculus. There we will illustrate the use of the dagger channel in basic techniques in machine learning, see Section 7.7.

Definition 6.6.1. Let $c \colon X \nrightarrow Y$ be a channel with a state $\omega \in \mathcal{D}(X)$ on its domain, such that the transformed / predicted state $c \gg\!= \omega \in \mathcal{D}(Y)$ has full support. The latter means that $\mathit{supp}(c \gg\!= \omega) = Y$, so that $(c \gg\!= \omega)(y) > 0$, for each $y \in Y$. Implicitly, the codomain Y is then a finite set — since supports are finite.

The dagger channel $c^\dagger_\omega \colon Y \nrightarrow X$ is defined as on y via backward inference using the point predicate $\mathbf{1}_y$, as in:

$$\begin{aligned} c^\dagger_\omega(y) \coloneqq \omega|_{c \gg\!= \mathbf{1}_y} &= \sum_{x \in X} \frac{\omega(x) \cdot c(x)(y)}{(c \gg\!= \omega)(y)} \big|x\big\rangle \\ &= \sum_{x \in X} \frac{\omega(x) \cdot c(x)(y)}{\sum_z \omega(z) \cdot c(z)(y)} \big|x\big\rangle. \end{aligned} \tag{6.6}$$

Thus, the dagger is obtained by updating the prior ω via the likelihood function $Y \to \mathcal{P}red(X)$, given by $y \mapsto c \gg\!= \mathbf{1}_y$, see Definition 4.3.1. We have already implicitly used the dagger of a channel in several situations.

Example 6.6.2. 1 In Example 6.3.2 we used a channel $c \colon \{H, T\} \nrightarrow \{W, B\}$ from sides of a coin to colors of balls in an urn, together with a fair coin $\mathit{unif} \in \mathcal{D}(\{H, T\})$. We had evidence of a white ball and wanted to know the updated coin distribution. The outcome that we computed can be described via a dagger channel, since:

$$c^\dagger_{\mathit{unif}}(W) = \mathit{unif}|_{c \gg\!= \mathbf{1}_W} = \tfrac{22}{67}|H\rangle + \tfrac{45}{67}|T\rangle.$$

The same redescription in terms of Bayesian inversion can be used for Examples 6.3.3 and 6.3.9, 6.3.10. In Example 6.3.5 we can use Bayesian inversion for the point evidence case, but not for the illustration with soft evidence, *i.e.* with fuzzy predicates. This matter will be investigated further in Section 6.7.

2 In Exercises 6.3.1 and 6.3.2 we have seen a channel $c \colon \{d, d^\perp\} \nrightarrow \{p, n\}$ that combines the sensitivity and specificity of a medical test, in a situation with a prior / prevalence of 1% for the disease. The associated Positive Prediction Value (PPV) and Negative Predication Value (NPV) can be expressed via a dagger as:

$$\begin{aligned} \mathit{PPV} &= c^\dagger_\omega(p)(d) = \omega|_{c \gg\!= \mathbf{1}_p}(d) = \tfrac{18}{117} \\ \mathit{NPV} &= c^\dagger_\omega(n)(d^\perp) = \omega|_{c \gg\!= \mathbf{1}_n}(d^\perp) = \tfrac{1881}{1883}, \end{aligned}$$

where $\omega = \frac{1}{100}|d\rangle + \frac{99}{100}|d^\perp\rangle$ is the prior distribution.

In Definition 6.3.1 we have carefully distinguished forward inference $c \gg\!= (\omega|_p)$ from backward inference $\omega|_{c \gg\!= q}$. Since Bayesian inversion turns channels around, the question arises whether it also turns inference around: can one express forward (resp. backward) inference along c in terms of backward (resp. forward) inference along the Bayesian inversion of c? The next result from [77] shows that this is indeed the case. It demonstrates that the directions of inference and of channels are intimately connected and it gives the channel-based framework internal consistency.

Theorem 6.6.3. *Let $c\colon X \nrightarrow Y$ be a channel with a state $\sigma \in \mathcal{D}(X)$ on its domain, such that the transformed state $\tau \coloneqq c \gg\!= \sigma$ has full support.*

1 *Given a factor q on Y, we can express backward inference as forward inference with the dagger:*

$$\sigma|_{c \mathrel{=\!\!\ll} q} = c_\sigma^\dagger \gg\!= (\tau|_q).$$

2 *Given a factor p on X, we can express forward inference as backward inference with the dagger:*

$$c \gg\!= (\sigma|_p) = \tau\Big|_{c_\sigma^\dagger \mathrel{=\!\!\ll} p}.$$

Proof. 1 For $x \in X$ we have:

$$\begin{aligned}
\Big(c_\sigma^\dagger \gg\!= (\tau|_q)\Big)(x) &= \sum_{y \in Y} c_\sigma^\dagger(y)(x) \cdot (c \gg\!= \sigma)|_q(y) \\
&= \sum_{y \in Y} \frac{c(x)(y) \cdot \sigma(x)}{(c \gg\!= \sigma)(y)} \cdot \frac{(c \gg\!= \sigma)(y) \cdot q(y)}{c \gg\!= \sigma \models q} \\
&= \frac{\sigma(x) \cdot \big(\sum_{y \in Y} c(x)(y) \cdot q(y)\big)}{\sigma \models c \mathrel{=\!\!\ll} q} \\
&= \frac{\sigma(x) \cdot (c \mathrel{=\!\!\ll} q)(x)}{\sigma \models c \mathrel{=\!\!\ll} q} \\
&= \sigma|_{c \mathrel{=\!\!\ll} q}(x).
\end{aligned}$$

2 Similarly, for $y \in Y$,

$$\begin{aligned}
\Big(\tau\Big|_{c_\sigma^\dagger \mathrel{=\!\!\ll} p}\Big)(y) &= \frac{(c \gg\!= \sigma)(y) \cdot (c_\sigma^\dagger \mathrel{=\!\!\ll} p)(y)}{c \gg\!= \sigma \models c_\sigma^\dagger \mathrel{=\!\!\ll} p} \\
&= \sum_{x \in X} \frac{(c \gg\!= \sigma)(y) \cdot c_\sigma^\dagger(y)(x) \cdot p(x)}{c_\sigma^\dagger \gg\!= (c \gg\!= \sigma) \models p} \\
&= \sum_{x \in X} \frac{(c \gg\!= \sigma)(y) \cdot \frac{\sigma(x) \cdot c(x)(y)}{(c \gg\!= \sigma)(y)} \cdot p(x)}{\sigma \models p} \\
&= \sum_{x \in X} c(x)(y) \cdot \frac{\sigma(x) \cdot p(x)}{\sigma \models p} \\
&= \sum_{x \in X} c(x)(y) \cdot \sigma|_p(x) \\
&= \Big(c \gg\!= (\sigma|_p)\Big)(y).
\end{aligned}$$

❑

We include two further results with basic properties of the dagger of a channel. More such properties are investigated in Section 7.8, where the dagger is put in a categorical perspective.

Proposition 6.6.4. *Let $c\colon X \multimap Y$ be a channel with a state $\sigma \in \mathcal{D}(X)$ such that $c \gg\!= \sigma$ has full support. Then:*

1 *Transforming the predicted state along the dagger channel yields the original state:*

$$c^{\dagger}_{\sigma} \gg\!= (c \gg\!= \sigma) = \sigma.$$

2 *The double dagger of a channel is the channel itself:*

$$\left(c^{\dagger}_{\sigma}\right)^{\dagger}_{c \gg\!= \sigma} = c.$$

Proof. 1 For $x \in X$ we have:

$$\begin{aligned}
\Big(c^{\dagger}_{\sigma} \gg\!= (c \gg\!= \sigma)\Big)(x) &= \sum_{y \in Y} c^{\dagger}_{\sigma}(y)(x) \cdot (c \gg\!= \sigma)(y) \\
&\overset{(6.6)}{=} \sum_{y \in Y} \frac{\sigma(x) \cdot c(x)(y)}{(c \gg\!= \sigma)(y)} \cdot (c \gg\!= \sigma)(y) \\
&= \sigma(x) \cdot \left(\sum_{y \in Y} c(x)(y)\right) \\
&= \sigma(x).
\end{aligned}$$

2 Again we unfold the definitions: for $x \in X$ and $y \in Y$,

$$\begin{aligned}
\left(c^{\dagger}_{\sigma}\right)^{\dagger}_{c \gg\!= \sigma}(x)(y) &\overset{(6.6)}{=} \frac{c^{\dagger}_{\sigma}(y)(x) \cdot (c \gg\!= \sigma)(y)}{\big(c^{\dagger}_{\sigma} \gg\!= (c \gg\!= \sigma)\big)(x)} \\
&= \frac{\sigma(x) \cdot c(x)(y) \cdot (c \gg\!= \sigma)(y)}{(c \gg\!= \sigma)(y) \cdot \sigma(x)} \quad \text{by (6.6) and item (1)} \\
&= c(x)(y).
\end{aligned}$$

❑

Next we show the close connection between channels, and their daggers, and joint states.

Theorem 6.6.5 (Disintegration)**.** *Consider the two projections $X \xleftarrow{\pi_1} X \times Y \xrightarrow{\pi_2} Y$ as deterministic channels. Let $\omega \in \mathcal{D}(X \times Y)$ be a joint state, whose marginals $\omega_1 := \omega[1, 0] = \pi_1 \gg\!= \omega \in \mathcal{D}(X)$ and $\omega_2 := \omega[0, 1] = \pi_2 \gg\!= \omega \in \mathcal{D}(Y)$ both have full support.*

Extract from ω the two channels:

$$c := \left(X \xrightarrow{(\pi_1)^{\dagger}_{\omega}} X \times Y \xrightarrow{\pi_2} Y\right) \qquad d := \left(Y \xrightarrow{(\pi_2)^{\dagger}_{\omega}} X \times Y \xrightarrow{\pi_1} X\right)$$

Then:

1 *The joint state ω is the graph of both these channels:*

$$\langle id, c\rangle \gg\!= \omega_1 = \omega = \langle d, id\rangle \gg\!= \omega_2.$$

2 The extracted channels are each other's daggers:

$$c^{\dagger}_{\omega_1} = d \qquad \text{and} \qquad d^{\dagger}_{\omega_2} = c.$$

This process of extracting channels from a joint state is called *disintegration*. It will be studied more systematically in the next chapter, see Section 7.6. The two channels $X \to Y$ and $Y \to X$ that we can extract in different directions from a joint state on $X \times Y$ are each other's daggers. In traditional probability this is expressed as equation:

$$P(A \mid B) \cdot P(B) = P(A, B) = P(B \mid A) \cdot P(A).$$

Proof. For both items we only do the first equations, since the second ones follow by symmetry.

1 For $x \in X$ and $y \in Y$,

$$\begin{aligned}
\Big(\langle id, c\rangle \gg\!= \omega_1\Big)(x, y) &= \omega_1(x) \cdot (\pi_2 \circ (\pi_1)^{\dagger}_{\omega})(x)(y) \\
&= \omega_1(x) \cdot \left(\sum_{z \in X} (\pi_1)^{\dagger}_{\omega}(x)(z, y) \right) \\
&\overset{(6.6)}{=} (\pi_1 \gg\!= \omega)(x) \cdot \left(\sum_{z \in X} \frac{\omega(z, y) \cdot \langle \pi_1 \rangle((z, y), x)}{(\pi_1 \gg\!= \omega)(x)} \right) \\
&= \omega(x, y).
\end{aligned}$$

2 Similarly,

$$\begin{aligned}
c^{\dagger}_{\omega_1}(y)(x) \overset{(6.6)}{=} \frac{\omega_1(x) \cdot c(x)(y)}{(c \gg\!= \omega_1)(y)} &= \frac{(\langle id, c\rangle \gg\!= \omega_1)(x, y)}{\omega_2(y)} \\
&= \frac{\omega(x, y)}{(\pi_2 \gg\!= \omega)(y)} \\
&= \sum_{z \in Y} (\pi_2)^{\dagger}_{\omega}(y)(x, z) \\
&= (\pi_1 \circ (\pi_2)^{\dagger}_{\omega})(y)(x) \\
&= d(y)(x). \qquad \square
\end{aligned}$$

We recall the Ewens draw-delete/add maps with the Ewens distributions from Section 3.9, in a situation:

$$\mathcal{E}(K) \underset{EDA(t)}{\overset{EDD}{\rightleftarrows}} \mathcal{E}(K+1) \qquad \text{with} \qquad ew[K](t) \in \mathcal{D}(\mathcal{E}(K)).$$

As announced there, these *EDD* and *EDA* maps are each other's daggers. We can now make this precise.

Proposition 6.6.6. *The Ewens draw-delete/add channels are each other's daggers, via the Ewens distributions: for each $K, t \in \mathbb{N}_{>0}$ one has:*

$$EDD^{\dagger}_{ew[K+1](t)} = EDA(t) \qquad \text{and} \qquad EDA(t)^{\dagger}_{ew[K](t)} = EDD.$$

Proof. Fix Ewens multisets $\varphi \in \mathcal{E}(K)$ and $\psi \in \mathcal{E}(K+1)$. A useful observation to start with is that the commuting diagrams in Corollary 3.9.12 give us the following validities of transformed point predicates.

$$\begin{aligned} ew[K+1](t) \models EDD \mathrel{=\!\!\ll} \mathbf{1}_{\varphi} &= (EDD \gg\!= ew[K+1](t))(\varphi) \\ &= ew[K](t)(\varphi) \\ ew[K](t) \models EDA(t) \mathrel{=\!\!\ll} \mathbf{1}_{\psi} &= (EDA(t) \gg\!= ew[K](t))(\psi) \\ &= ew[K+1](t)(\psi). \end{aligned}$$

For the first equation in the proposition we mimick the proof of Corollary 3.9.12 and follow the reasoning there:

$$\begin{aligned} & EDD^{\dagger}_{ew[K+1](t)}(\varphi) \\ &\overset{(6.6)}{=} \sum_{\psi \in \mathcal{E}(K+1)} \frac{ew[K+1](t)(\psi) \cdot (EDD \mathrel{=\!\!\ll} \mathbf{1}_{\varphi})(\psi)}{ew[K+1](t) \models EDD \mathrel{=\!\!\ll} \mathbf{1}_{\varphi}} \big|\psi\big\rangle \\ &= \sum_{\psi \in \mathcal{E}(K+1)} \frac{ew[K+1](t)(\psi) \cdot EDD(\psi)(\varphi)}{ew[K](t)(\varphi)} \big|\psi\big\rangle \\ &= \frac{\varphi(1)+1}{K+1} \cdot \frac{ew[K+1](t)(\varphi + 1|1\rangle)}{ew[K](t)(\varphi)} \big|\varphi + 1|1\rangle\big\rangle \\ &\qquad + \sum_{1 \leq k \leq K} \frac{(\varphi(k+1)+1) \cdot (k+1)}{K+1} \\ &\qquad\qquad \cdot \frac{ew[K+1](t)(\varphi - 1|k\rangle + 1|k+1\rangle)}{ew[K](t)(\varphi)} \big|\varphi - 1|k\rangle + 1|k+1\rangle\big\rangle \\ &= \frac{t}{K+t}\big|\varphi + 1|1\rangle\big\rangle + \sum_{1 \leq k \leq K} \frac{\varphi(k) \cdot k}{K+t} \big|\varphi - 1|k\rangle + 1|k+1\rangle\big\rangle \\ &= EDA(t)(\varphi). \end{aligned}$$

Similarly, using the approach in the proof of Theorem 3.9.10,

$$\begin{aligned}
&EDA(t)^{\dagger}_{ew[K](t)}(\psi) \\
&\overset{(6.6)}{=} \sum_{\varphi\in\mathcal{E}(K)} \frac{ew[K](t)(\varphi)\cdot(EDA(t) =\!\!\ll \mathbf{1}_\psi)(\varphi)}{ew[K](t) \models EDA(t) =\!\!\ll \mathbf{1}_\psi} \big|\varphi\big\rangle \\
&= \sum_{\varphi\in\mathcal{E}(K)} \frac{ew[K](t)(\varphi)\cdot EDA(t)(\psi)(\varphi)}{ew[K+1](t)(\psi)} \big|\varphi\big\rangle \\
&= \frac{t}{K+t}\cdot\frac{ew[K](t)(\psi-1|1\rangle)}{ew[K+1](t)(\psi)}\big|\psi-1|1\rangle\big\rangle \qquad (\text{if } \psi(1)>0) \\
&\quad + \sum_{1\leq k\leq K} \frac{(\psi(k)+1)\cdot k}{K+t} \\
&\qquad\qquad \cdot \frac{ew[K](t)(\psi+1|k\rangle - 1|k+1\rangle)}{ew[K+1](t)(\psi)}\big|\psi+1|k\rangle - 1|k+1\rangle\big\rangle \\
&= \frac{t}{K+t}\cdot\frac{\prod_{1\leq\ell\leq K}\binom{t}{\ell}^{(\psi-1|1\rangle)(\ell)}}{(\psi-1|1\rangle)\binom{t}{K}}\cdot\frac{\psi\binom{t}{K+1}}{\prod_{1\leq\ell\leq K+1}\binom{t}{\ell}^{\psi(\ell)}}\big|\psi-1|1\rangle\big\rangle \\
&\quad + \sum_{1\leq k\leq K} \frac{(\psi(k)+1)\cdot k}{K+t}\cdot\frac{\prod_{1\leq\ell\leq K}\binom{t}{\ell}^{(\psi+1|k\rangle-1|k+1\rangle)(\ell)}}{(\psi+1|k\rangle-1|k+1\rangle)\binom{t}{K}} \\
&\qquad\qquad \cdot\frac{\psi\binom{t}{K+1}}{\prod_{1\leq\ell\leq K+1}\binom{t}{\ell}^{\psi(\ell)}}\big|\psi+1|k\rangle-1|k+1\rangle\big\rangle \\
&= \frac{t}{K+t}\cdot\psi(1)\cdot\frac{K+t}{K+1}\cdot\frac{1}{t}\big|\psi-1|1\rangle\big\rangle \qquad (\text{if } \psi(1)>0,\ \text{so } \psi(K+1)=0) \\
&\quad + \sum_{1\leq k\leq K}\frac{(\psi(k)+1)\cdot k}{K+t}\cdot\frac{\psi(k+1)}{\psi(k)+1}\cdot\frac{K+t}{K+1}\cdot\frac{t/k}{t/(k+1)}\big|\psi+1|k\rangle-1|k+1\rangle\big\rangle \\
&= \frac{\psi(1)}{K+1}\big|\psi-1|1\rangle\big\rangle + \sum_{2\leq k\leq K+1}\frac{\psi(k)\cdot k}{K+1}\big|\psi+1|k-1\rangle-1|k\rangle\big\rangle \\
&= EDD(\varphi). \qquad \blacksquare
\end{aligned}$$

Exercises

6.6.1 Let $\sigma,\tau\in\mathcal{D}(X)$ be distributions with $supp(\tau)\subseteq supp(\sigma)$. Show that for the identity channel id one has:

$$id^{\dagger}_{\sigma} =\!\!\gg \tau = \tau.$$

6.6.2 Let $c\colon X \nrightarrow Y$ be a channel with state $\omega\in\mathcal{D}(X)$, where $c =\!\!\gg \omega$ has full support. Consider the following distribution of distributions:

$$\Omega \coloneqq \sum_{y\in Y}(c =\!\!\gg \omega)(y)\big|c^{\dagger}_{\omega}(y)\big\rangle \in \mathcal{D}(\mathcal{D}(X)).$$

Recall the probabilistic 'flatten' operation from Section 2.2 and show:

$$\mathit{flat}(\Omega) = \omega.$$

This says that Ω is 'Bayes plausible' in the terminology of [97] on Bayesian persuasion. The construction of Ω is one part of the bijective correspondence in Exercise 7.6.7.

6.6.3 For $\omega \in \mathcal{D}(X)$, $c \colon X \nrightarrow Y$, $p \in \mathit{Obs}(X)$ and $q \in \mathit{Obs}(Y)$ prove:

$$c \gg\!= \omega \models (c^{\dagger}_{\omega} \mathrel{=\!\!\ll} p) \mathbin{\&} q \;=\; \omega \models p \mathbin{\&} (c \mathrel{=\!\!\ll} q).$$

This is a reformulation of [24, Thm. 6]. If we ignore the validity signs $\models$ and the states before them, then we can recognise in this equation the familiar 'adjointness property' of daggers (adjoint transposes) in the theory of Hilbert spaces: $\langle A^{\dagger}(x) \mid y \rangle = \langle x \mid A(y) \rangle$.

6.6.4 Consider the draw-delete channel $DD \colon \mathcal{N}[K+1](X) \nrightarrow \mathcal{N}[K](X)$ from Definition 2.2.4. Let X be a finite set, say with n elements. Consider the the uniform state *unif* on $\mathcal{N}[K+1](X)$, see Exercise 2.3.7. Show that DD's dagger, with respect to *unif*, can be described on $\varphi \in \mathcal{N}[K](X)$ as:

$$DD^{\dagger}_{\mathit{unif}}(\varphi) = \sum_{x \in X} \frac{\varphi(x) + 1}{K + n} \big| \varphi + 1 | x \big\rangle\!\big\rangle.$$

This dagger differs from the draw-add map $DA \colon \mathcal{N}[K](X) \nrightarrow \mathcal{N}[K+1](X)$.

6.6.5 Let $\omega \in \mathcal{D}(X \times Y)$ be a joint state, whose two marginals $\omega_1 \coloneqq \omega[1,0] \in \mathcal{D}(X)$ and $\omega_2 \coloneqq \omega[0,1] \in \mathcal{D}(Y)$ have full support. Let $p \in \mathit{Pred}(X)$ and $q \in \mathit{Pred}(Y)$ be arbitrary predicates.

Prove that there are predicates q_1 on X and p_2 on Y such that:

$$\omega_1 \models p \mathbin{\&} q_1 \;=\; \omega \models p \otimes q \;=\; \omega_2 \models p_1 \mathbin{\&} q.$$

This is a discrete version of [131, Prop. 6.7].

Hint: Disintegrate!

6.6.6 Let $c \colon X \nrightarrow Y$ and $d \colon X \nrightarrow Z$ be channels with a state $\omega \in \mathcal{D}(X)$ on their domain such that both predicted states $c \gg\!= \omega$ and $d \gg\!= \omega$ have full support. Show that:

$$\begin{aligned} \langle c, d \rangle^{\dagger}_{\omega}(y, z) &= \omega|_{(c \mathrel{=\!\!\ll} \mathbf{1}_y) \mathbin{\&} (d \mathrel{=\!\!\ll} \mathbf{1}_z)} \;=\; d^{\dagger}_{c^{\dagger}_{\omega}(y)}(z) \\ &= \omega|_{(d \mathrel{=\!\!\ll} \mathbf{1}_z) \mathbin{\&} (c \mathrel{=\!\!\ll} \mathbf{1}_y)} \;=\; c^{\dagger}_{c^{\dagger}_{\omega}(z)}(y). \end{aligned}$$

6.6.7 Recall from Exercise 2.2.11 what it means for $c \colon X \nrightarrow Y$ to be a bi-channel. Check that $c^{\dagger} \colon Y \nrightarrow Y$, as defined there, is $c^{\dagger}_{\mathit{unif}}$, where *unif* is the uniform distribution on X (assuming that X is finite).

6.7 Pearl's and Jeffrey's update rules

We have seen that backward inference is a useful reasoning technique, in a situation where we have a state $\sigma \in \mathcal{D}(X)$, on the domain X of a channel $c \colon X \multimap Y$, that we wish to update in the light of evidence on the codomain Y of the channel. The evidence is given in the form of a predicate (or factor) on Y. This backward inference is also called Pearl's update rule.

It turns out that there is an alternative update mechanism in this situation, where the evidence is not given by a predicate on the codomain of the channel, but by state. This alternative was introduced by Jeffrey [90], see also [19, 33, 39, 160], or [17] for a recent application in physics. Jeffrey's rule can be formulated conveniently (see [77]) in terms of the Bayesian inverse (dagger) $c^\dagger$ of the channel — as introduced in the previous section.

As we shall see below, the update rules of Pearl and Jeffrey can give quite different outcomes. Hence the question arises: when to use which rule? What is the difference? There is a clear difference, which can be summarised as follows: Pearl's rule increases validity and Jeffrey's rule decreases divergence. This will be made technically precise, in Theorem 6.7.4 below. The proof in Pearl's case is easy, but the proof in Jeffrey's case is remarkably difficult. We copy it from [79] and include it in a separate subsection.

In general terms, one can learn by reinforcing what goes well, or by steering away from what goes wrong. In the first case one improves a positive evaluation and in the second case one reduces a negative outcome, as a form of correction. Thus, Jeffrey's update rule is a correction mechanism, for correcting errors. As such it is used in 'predictive coding', a theory in cognitive science that views the human brain as a Bayesian prediction engine, see [143, 51, 64, 23].

This section introduced Jeffrey's update rule, especially in contrast to Pearl's rule, that is, to backward inference.

Definition 6.7.1. Let $c \colon X \multimap Y$ be channel with a prior state $\sigma \in \mathcal{D}(X)$.

1 *Pearl's rule* is backward inference: it involves updating of the prior σ with evidence in the form of a *factor* $q \in \mathit{Fact}(Y)$ to the posterior:

$$\sigma|_{c \gg q} \in \mathcal{D}(X).$$

2 *Jeffrey's rule* involves update of the prior σ with evidence in form of a *state* $\tau \in \mathcal{D}(Y)$ to the posterior:

$$c^\dagger_\sigma \gg\!= \tau \in \mathcal{D}(X).$$

We shall illustrate the difference between Pearl's and Jeffreys's rules in two examples.

Example 6.7.2. Consider the situation from Exercise 6.3.1 with set $D = \{d, d^\perp\}$ for disease (or not) and $T = \{p, n\}$ for a positive or negative test outcome, with a test channel $c\colon D \nrightarrow T$ given by:

$$c(d) = \tfrac{9}{10}|p\rangle + \tfrac{1}{10}|n\rangle \qquad c(d^\perp) = \tfrac{1}{20}|p\rangle + \tfrac{19}{20}|n\rangle.$$

The test thus has a sensitivity of 90% and a specificity of 95%. We assume a prevalence of 10% via a prior state $\omega = \frac{1}{10}|d\rangle + \frac{9}{10}|d^\perp\rangle$.

The test is performed, under unfavourable circumstances like bad light, and we are only 80% sure that the test is positive. With Pearl's update rule we thus use as evidence predicate $q = \frac{4}{5} \cdot \mathbf{1}_p + \frac{1}{5} \cdot \mathbf{1}_n$. It gives as posterior disease probability:

$$\omega|_{c \mathrel{\scriptscriptstyle\rightarrowtail} q} = \tfrac{74}{281}|d\rangle + \tfrac{207}{281}|\,d\rangle \approx 0.26|d\rangle + 0.73|\,d\rangle.$$

This gives a disease likelihood of 26%.

When we decide to use Jeffrey's rule we translate the 80% certainty of a positive test into a state $\tau = \frac{4}{5}|p\rangle + \frac{1}{5}|n\rangle$. Then we compute:

$$\begin{aligned} c^\dagger_\sigma \gg\!= \tau &= \tfrac{4}{5} \cdot c^\dagger_\sigma(p) + \tfrac{1}{5} \cdot c^\dagger_\sigma(n) \\ &\overset{(6.6)}{=} \tfrac{4}{5} \cdot \omega|_{c \mathrel{\scriptscriptstyle\rightarrowtail} \mathbf{1}_p} + \tfrac{1}{5} \cdot \omega|_{c \mathrel{\scriptscriptstyle\rightarrowtail} \mathbf{1}_n} \\ &= \tfrac{4}{5} \cdot \Big(\tfrac{2}{3}|d\rangle + \tfrac{1}{3}|\,d\rangle\Big) + + \tfrac{4}{5} \cdot \Big(\tfrac{2}{173}|d\rangle + \tfrac{171}{173}|\,d\rangle\Big) \\ &= \tfrac{278}{519}|d\rangle + \tfrac{241}{519}|\,d\rangle \\ &\approx 0.54|d\rangle + 0.46|\,d\rangle. \end{aligned}$$

The disease likelihood is now 54%, more than twice as high as with Pearl's update rule. This is a serious difference, which may have serious consequences. Should we start asking our doctors: does your computer use Pearl's or Jeffrey's rule to calculate likelihoods, and come to an advice about medical treatment?

We also review an earlier illustration, now with Jeffrey's approach.

Example 6.7.3. Recall the situation of Example 2.2.2, with a teacher predicting the performance of pupils, depending on the teacher's mood. The evidence q from Example 6.3.4 is used for Pearl's update. It can be translated into a state τ on the set G of grades:

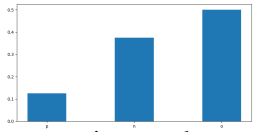
$$\tau = \tfrac{1}{10}|1\rangle + \tfrac{3}{10}|2\rangle + \tfrac{3}{10}|3\rangle + \tfrac{2}{10}|4\rangle + \tfrac{1}{10}|5\rangle.$$

There is an a priori divergence $D_{KL}(\tau, c \gg\!= \sigma) \approx 1.336$. With some effort one can prove that the Jeffrey-update of σ is:

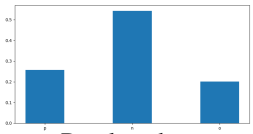
$$\begin{aligned} \sigma' = c^\dagger_\sigma \gg\!= \tau &= \tfrac{972795}{3913520}|p\rangle + \tfrac{1966737}{3913520}|n\rangle + \tfrac{973988}{3913520}|o\rangle \\ &\approx 0.2486|p\rangle + 0.5025|n\rangle + 0.2489|o\rangle. \end{aligned}$$

The divergence has now dropped to: $D_{KL}(\tau, c \gg\!= \sigma') \approx 1.087$.

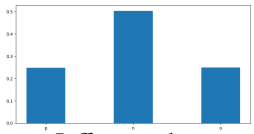
In the end it is interesting to compare the orinal (prior) mood with its Pearl- and Jeffrey-updates. In both cases we see that after the bad marks of the pupils the teacher has become less optimistic.



prior mood



Pearl-update



Jeffrey-update

The prior mood is reproduced from Example 2.2.2 for easy comparison. The Pearl and Jeffrey updates differ only slightly in this case.

The following theorem (from [79]) captures the essences of the update rules of Pearl and Jeffrey: validity increase or divergence decrease.

Theorem 6.7.4. *Consider the situation in Definition 6.7.1, with a prior state $\sigma \in \mathcal{D}(X)$ and a channel $c\colon X \nrightarrow Y$.*

1 *For a factor q on Y Pearl's update rule gives an increase of validity:*

$$c \gg\!= \sigma_P \models q \;\geq\; c \gg\!= \sigma \models q \quad \text{for the updated state} \quad \sigma_P = \sigma|_{c \ll q}.$$

2 *For an evidence state $\tau \in \mathcal{D}(Y)$ Jeffrey's update rule gives a decrease of Kullback-Leibler divergence:*

$$D_{KL}(\tau, c \gg\!= \sigma_J) \leq D_{KL}(\tau, c \gg\!= \sigma) \qquad \text{for} \qquad \sigma_J = c_\sigma^\dagger \gg\!= \tau.$$

In this situation we assume that the predicted state $c \gg\!= \sigma$ as fulll support, so that the dagger $c_\sigma^\dagger$ is well-defined.

In this sitation $c \gg\!= \sigma$ is the predicted state, where we evaluate the evidence.

- In Pearl's case we look at the validity $c \gg\!= \sigma \models q$ of the evidence q in this predicted state. The above theorem tells that by switching to the Pearl-update σ_P that we get a higher validity $c \gg\!= \sigma_P \models q$.
- In Jeffrey's case we look at the divergence / mismatch $D_{KL}(\tau, c \gg\!= \sigma)$ between the evidence state τ and the predicted state $c \gg\!= \sigma$. By changing to the

Jeffrey-update σ_J we get a lower divergence $D_{KL}(\tau, c \gg\!= \sigma_J)$. Thus, via Jeffrey's rule one reduces 'prediction errors', in the terminology of predictive coding theory.

Proof. 1 Via Proposition 4.3.3 we can turn $\gg\!=$ on the left of $\models$ into $=\!\ll$ on the right and vice-versa, so that we can use the validity increase of Theorem 6.1.5:

$$\begin{aligned} c \gg\!= (\sigma|_{c =\!\ll q}) \models q &= \sigma|_{c =\!\ll q} \models c =\!\ll q \\ &\geq \sigma \models c =\!\ll q \\ &= c \gg\!= \sigma \models q. \end{aligned}$$

2 The proof is non-trivial and relegated to Subsection 6.7.1. ❑

The fact that Jeffrey's rule involves correction of prediction errors, as stated in Theorem 6.7.4 (2), has led to the view that Jeffrey's update rule should be used in situations where one is confronted with 'surprises' [41] or with 'unanticipated knowledge' [40]. Here we include an example from [41] (also used in [77]) that involves such error correction after a 'surprise'.

Example 6.7.5. Ann must decide about hiring Bob, whose characteristics are described in terms of a combination of competence (c or $c^\perp$) and experience (e or $e^\perp$). The prior, based on experience with many earlier candidates, is a joint distribution on the product space $C \times E$, for $C = \{c, c^\perp\}$ and $E = \{e, e^\perp\}$, given as:

$$\omega = \tfrac{4}{10}|c, e\rangle + \tfrac{1}{10}|c, e^\perp\rangle + \tfrac{1}{10}|c^\perp, e\rangle + \tfrac{4}{10}|c^\perp, e^\perp\rangle.$$

The first marginal of ω is the uniform distribution $\frac{1}{2}|c\rangle + \frac{1}{2}|c^\perp\rangle$. It is the neutral base rate for Bob's competence.

We use the two projection functions $C \xleftarrow{\pi_1} C \times E \xrightarrow{\pi_2} E$ as deterministic channels along which we reason, using both Pearl's and Jeffrey's rules.

When Ann would learn that Bob has relevant work experience, given by point evidence $\mathbf{1}_e$, her strategy is to factor this in via Pearl's rule / backward inference: this gives as posterior $\omega|_{\pi_2 =\!\ll \mathbf{1}_e} = \omega|_{\mathbf{1} \otimes \mathbf{1}_e}$, whose first marginal is $\frac{4}{5}|c\rangle + \frac{1}{5}|c^\perp\rangle$. It is then more likely that Bob is competent.

Ann reads Bob's letter to find out if he actually has relevant experience. We quote from [41]:

> Bob's answer reveals right from the beginning that his written English is poor. Ann notices this even before figuring out what Bob says about his work experience. In response to this unforeseen learnt input, Ann lowers her probability that Bob is competent from $\frac{1}{2}$ to $\frac{1}{8}$. It is natural to model this as an instance of Jeffrey revision.

Bob's poor English is a new state of affairs: a surprise that causes Ann to switch to error reduction mode, via Jeffrey's rule. Bob's poor command of the English language translates into a competence state $\rho = \frac{1}{8}|c\rangle + \frac{7}{8}|c^\perp\rangle$. Ann wants to adapt to this new surprising situation, so she uses Jeffrey's rule, giving a new joint state:

$$\omega' = (\pi_1)^\dagger_\omega \gg\!= \rho = \tfrac{1}{10}|c, e\rangle + \tfrac{1}{40}|c, e^\perp\rangle + \tfrac{7}{40}|c^\perp, e\rangle + \tfrac{7}{10}|c^\perp, e^\perp\rangle.$$

If the letter now tells that Bob has work experience, Ann will factor this in, in this new situation ω', giving, like above, via Pearl's rule followed by marginalisation,

$$\omega'|_{\pi_2 \prec\!\ll \mathbf{1}_e}[1, 0] = \tfrac{4}{11}|c\rangle + \tfrac{7}{11}|c^\perp\rangle.$$

The likelihood of Bob being competent is now lower than in the prior state, since $\frac{4}{11} < \frac{1}{2}$. This example reconstructs the illustration from [41] in channel-based form, with the associated formulations of Pearl's and Jeffrey's rules from Definition 6.7.1, and produces exactly the same outcomes as in *loc. cit.*

We briefly discuss some further commonalities and differences between the rules of Pearl and Jeffrey. We have seen some of these results before, but now they are put in the context of Pearl's and Jeffrey's update rule.

Proposition 6.7.6. *Let $c\colon X \multimap Y$ be a channel with a prior state $\sigma \in \mathcal{D}(X)$.*

1 *Pearl's rule and Jeffrey's rule agree on point predicate / states: for $y \in Y$, with associated point predicate $\mathbf{1}_y$ and point state $1|y\rangle$, one has:*

$$\sigma|_{c \prec\!\ll \mathbf{1}_y} = c^\dagger_\sigma(y) = c^\dagger_\sigma \gg\!= 1|y\rangle.$$

2 *Pearl's updating with a constant predicate (no information) does not change the prior state ω:*

$$\sigma|_{c \prec\!\ll r\cdot\mathbf{1}} = \sigma, \qquad \text{for } r > 0.$$

Jeffrey's update does not change anything when we update with what we already predict:

$$c^\dagger_\sigma \gg\!= (c \gg\!= \sigma) = \sigma.$$

This is in essence the law of total probability, see Exercise 6.7.6 below.

3 *For a factor $q \in \mathsf{Fact}(Y)$ we can update the state σ according to Pearl, and also the channel c so that the updated predicted state is predicted:*

$$c|_q \gg\!= (\sigma|_{c \prec\!\ll q}) = (c \gg\!= \sigma)|_q.$$

In Jeffrey's case, with an evidence state $\tau \in \mathcal{D}(Y)$, we can update both the

state and the channel, via a double-dagger, so that the evidence state τ is predicted:

$$\left(c^{\dagger}_{\sigma}\right)^{\dagger}_{\tau} \gg\!= \left(c^{\dagger}_{\sigma} \gg\!= \tau\right) = \tau.$$

4 *Multiple updates with Pearl's rule commute, but multiple updates with Jeffrey's rule do not commute.*

Proof. These items follow from earlier results.

1 Directly by Definition 6.7.1.
2 The first equation follows from Lemma 6.1.6 (1) and (4); the second one is Proposition 6.6.4 (1).
3 The first claim is Theorem 6.3.11 and the second one is an instance of Proposition 6.6.4 (2).
4 By Lemma 6.1.6 (3) we have:

$$\sigma|_{c \ll q_1}|_{c \ll q_2} = \sigma|_{(c \ll q_1)\&(c \ll q_2)} = \sigma|_{c \ll q_2}|_{c \ll q_1}.$$

However, in general, for evidence states $\tau_1, \tau_2 \in \mathcal{D}(Y)$,

$$c^{\dagger}_{c^{\dagger}_{\sigma} \gg\!= \tau_1} \gg\!= \tau_2 \neq c^{\dagger}_{c^{\dagger}_{\sigma} \gg\!= \tau_2} \gg\!= \tau_1.$$

Exercise 6.7.5 contains a concrete example. ❑

Item (2) present different views on 'learning', where learning is now used in an informal sense: according to Pearl's rule you learning nothing when you get no information; but according to Jeffrey you learn nothing when you are presented with what you already know. Both interpretations make sense.

Probabilistic updating may be used as a model for what is called *priming* in cognitive science, see *e.g.* [60, 64]. It is well-known that the human mind is sensitive to the order in which information is processed, that is, to the order of priming/updating. Thus, the last item (4) above suggests that Jeffrey's rule might be more appropriate in such a setting. This strengthens the view in predictive coding that the human mind learns from error correction, as in Jeffrey's update rule.

There are translations back-and-forth between Pearl's and Jeffrey's rules, due to [19]; they are translated here to the current setting.

Proposition 6.7.7. *Let $c\colon X \dashrightarrow Y$ be a channel with a prior state $\sigma \in \mathcal{D}(X)$ on its domain, such that the predicted state $c \gg\!= \sigma$ has full support.*

1 *Pearl's updating can be expressed as Jeffrey's update, by turning a factor $q\colon Y \to \mathbb{R}_{\geq 0}$ into a state $\rho|_q \in \mathcal{D}(Y)$, so that:*

$$c^{\dagger}_{\sigma} \gg\!= \left((c \gg\!= \sigma)\big|_q\right) = \sigma|_{c \ll q}.$$

2 *Jeffrey's updating can also be expressed as Pearl's updating: for an evidence state* $\tau \in \mathcal{D}(Y)$ *write* $\tau/_{(c \gg= \sigma)}$ *for the factor* $y \mapsto \frac{\tau(y)}{(c \gg= \sigma)(y)}$*; then:*

$$c^\dagger_\sigma \gg= \tau \;=\; \sigma|_{c =\!\!\ll \tau/_{(c \gg= \sigma)}}.$$

Proof. The first item is exactly Theorem 6.6.3 (1). For the second item we first note that $(c \gg= \sigma) \models \tau/_{(c \gg= \sigma)} = 1$, since $(c \gg= \sigma)$ is a state:

$$(c \gg= \sigma) \models \tau/_{(c \gg= \sigma)} \;=\; \sum_{y \in Y} (c \gg= \sigma)(y) \cdot \frac{\tau(y)}{(c \gg= \sigma)(y)} \;=\; \sum_{y \in Y} \tau(y) \;=\; 1. \tag{6.7}$$

But then, for $x \in X$,

$$\begin{aligned}
(c^\dagger_\sigma \gg= \tau)(x) &= \sum_{y \in Y} \tau(y) \cdot c^\dagger_\sigma(y)(x) \\
&\overset{(6.6)}{=} \sum_{y \in Y} \tau(y) \cdot \frac{\sigma(x) \cdot c(x)(y)}{(c \gg= \sigma)(y)} \\
&= \sigma(x) \cdot \sum_{y \in Y} c(x)(y) \cdot \frac{\tau(y)}{(c \gg= \sigma)(y)} \\
&= \frac{\sigma(x) \cdot (c =\!\!\ll \tau/_{(c \gg= \sigma)})(x)}{c \gg= \sigma \models \tau/_{(c \gg= \sigma)}} \qquad \text{as just shown} \\
&= \frac{\sigma(x) \cdot (c =\!\!\ll \tau/_{(c \gg= \sigma)})(x)}{\sigma \models c =\!\!\ll \tau/_{(c \gg= \sigma)}} \\
&= \sigma|_{c =\!\!\ll \tau/_{(c \gg= \sigma)}}(x).
\end{aligned}$$

❑

Combination of Theorem 6.7.4 and Proposition 6.7.7 tells us how Pearl's rule can also lead to a decrease of divergence. The situation is quite subtle and will be discussed further in the subsequent remarks.

Corollary 6.7.8. *Let* $c \colon X \multimap Y$ *be a channel with a prior state* $\sigma \in \mathcal{D}(X)$ *and a factor* q *on* Y*. Then:*

$$D_{KL}\Big((c \gg= \sigma)|_q,\; c \gg= (\sigma|_{c =\!\!\ll q})\Big) \;\leq\; D_{KL}\Big((c \gg= \sigma)|_q,\; c \gg= \sigma\Big).$$

This result can be interpreted as follows. With my prior state σ I can predict $c \gg= \sigma$. I can update this prediction with evidence q to $(c \gg= \sigma)|_q$. The divergence between this update and my prediction is bigger than the divergence between the update and the prediction from the Pearl/Bayes posterior $\sigma|_{c =\!\!\ll q}$. Thus, the posterior gives a correction.

Proof. Take $\tau = (c \gg= \sigma)|_q$. Theorem 6.7.4 (2) says that:

$$D_{KL}\Big(\tau, c \gg= (c^\dagger_\sigma \gg= \tau)\Big) \;\leq\; D_{KL}\Big(\tau, c \gg= \sigma\Big).$$

But Proposition 6.7.7 (1) tells that $c^\dagger_\sigma \gg= \tau = \sigma|_{c =\!\!\ll q}$. ❑

Remarks 6.7.9. 1 In the above corrolary we use that Pearl's rule can be expressed as Jeffrey's, see point (1) in Proposition 6.7.7. The other way around, Jeffrey's rule is expressed via Pearl's in point (2). This leads to a validity increase property for Jeffrey's rule: let $c\colon X \nrightarrow Y$ be a channel with states $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$. Assuming that $c \gg\!= \sigma$ has full support we can form the factor $q = \tau/(c \gg\!= \sigma)$. We then get an inequality:

$$c \gg\!= (c^{\dagger}_{\sigma} \gg\!= \tau) \models q \;=\; c \gg\!= (\sigma|_{c \ll q} \models q \;\geq\; c \gg\!= \sigma \models q \overset{(6.7)}{=} 1.$$

This is not very useful.

2 We have emphasised the message "Pearl increases validity" and "Jeffrey decreases divergence". But Corollary 6.7.8 seems to nuance this message, since it describes a divergence decrease for Pearl too, and the previous item gives a validity increase for Jeffrey. What is going on? We try to clarify the situation via an example, demonstrating that the validity increase of Pearl's rule fails for Jeffrey's rule and that the divergence decrease of Jeffrey's rule fails for Pearl's rule.

Take sets $X = \{0, 1\}$ and $Y = \{a, b, c\}$ with uniform prior $\sigma = \frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle \in \mathcal{D}(X)$. We use the channel $c\colon X \nrightarrow Y$ given by:

$$c(0) = \tfrac{1}{9}|a\rangle + \tfrac{2}{3}|b\rangle + \tfrac{2}{9}|c\rangle \quad \text{and} \quad c(1) = \tfrac{7}{25}|a\rangle + \tfrac{7}{25}|b\rangle + \tfrac{11}{25}|c\rangle.$$

The predicted state is then $c \gg\!= \sigma = \frac{44}{225}|a\rangle + \frac{71}{150}|b\rangle + \frac{149}{450}|c\rangle$. We use as 'equal' evidence predicate and state:

$$q = \tfrac{1}{2} \cdot \mathbf{1}_a + \tfrac{1}{3} \cdot \mathbf{1}_b + \tfrac{1}{6} \cdot \mathbf{1}_c \qquad \text{and} \qquad \tau = \tfrac{1}{2}|a\rangle + \tfrac{1}{3}|b\rangle + \tfrac{1}{6}|c\rangle.$$

We then get the following updates, according to Pearl and Jeffrey, respectively:

$$\begin{aligned} \sigma_P &= \sigma|_{c \ll q} \\ &= \tfrac{425}{839}|0\rangle + \tfrac{414}{839}|1\rangle \\ &\approx 0.5066|0\rangle + 0.4934|1\rangle \end{aligned} \qquad \text{and} \qquad \begin{aligned} \sigma_J &= c^{\dagger}_{\sigma} \gg\!= \tau \\ &= \tfrac{805675}{1861904}|0\rangle + \tfrac{1056229}{1861904}|1\rangle \\ &\approx 0.4327|0\rangle + 0.5673|1\rangle. \end{aligned}$$

The validities are summarised in the following table.

description	**formula**	**value**
prior validity	$c \gg\!= \sigma \models q$	0.31074
after Pearl	$c \gg\!= \sigma_P \models q$	0.31079
after Jeffrey	$c \gg\!= \sigma_J \models q$	0.31019

The differences are small, but relevant. Pearl's updating increases validity, as Theorem 6.7.4 (1) dictates, but Jeffrey's updating does not, in this example.

The divergences in this example are as follows.

description	formula	value
prior divergence	$D_{KL}(\tau, c \gg\!= \sigma)$	0.238
after Pearl	$D_{KL}(\tau, c \gg\!= \sigma_P)$	0.240
after Jeffrey	$D_{KL}(\tau, c \gg\!= \sigma_J)$	0.221

We see that Jeffrey's updating decreases divergence, in line with Theorem 6.7.4 (2), but Pearl's updating does not. Is there a contradiction with Corollary 6.7.8, which does involve a divergence decrease? No, since there the state τ has a very particular shape, namely $(c \gg\!= \sigma)|_q$. We conclude that for that particular state Pearl's rule gives a divergence decrease, but there is no such decrease in general.

Thus, we conclude that, in general, validity increase is exclusive for Pearl's rule, and divergence decrease is exclusive for Jeffrey's rule.

6.7.1 Proof of the Jeffrey's divergence reduction

We have skipped a proof of the divergence reduction in Jeffrey's update rule, in Theorem 6.7.4 (2). Here we include a non-trivial proof, copied from [79]. It uses the following result, which we shall call the 'weigthed update inequality'.

Proposition 6.7.10. *Let $\omega \in \mathcal{D}(X)$ be a state with predicates $p_1, \ldots, p_n \in \mathrm{Pred}(X)$ forming a 'test', so that $\bigvee_i p_i = \mathbf{1}$. We assume $\omega \models p_i \neq 0$, for each i. For all numbers $r_1, \ldots, r_n \in \mathbb{R}_{\geq 0}$ with $\sum_i r_i = 1$, one has:*

$$\sum_i \frac{r_i \cdot (\omega \models p_i)}{\sum_j r_j \cdot (\omega|_{p_j} \models p_i)} \;\leq\; 1. \tag{6.8}$$

This weighted update inequality follows from [50, Thm. 4.1]. The proof is non-trivial, but involves some interesting properties of matrices of validities. We skip this proof now and first show how it is used in a proof of the divergence reduction of Jeffrey's rule. A binary version, for $n = 2$, of this result has appeared already in Exercise 6.1.12.

We use the name 'weigthed update', since the denominator in (6.8) can also be written as the validity of the predicate p_i in the weighted sum of updates:

$$\textstyle\sum_j r_j \cdot \big((\omega|_{p_j} \models p_i\big) = \big(\sum_j r_j \cdot \omega|_{p_j}\big) \models p_i.$$

Proof. [of Theorem 6.7.4 (2)] We reason as follows:

$$\begin{aligned}
&D_{KL}\Big(\tau, c \gg\!= (c^\dagger_\sigma \gg\!= \tau)\Big) \;-\; D_{KL}\Big(\tau, c \gg\!= \sigma\Big) \\
&= \sum\nolimits_y \tau(y) \cdot \log\left(\frac{\tau(y)}{(c \gg\!= (c^\dagger_\sigma \gg\!= \tau))(y)}\right) \;-\; \sum\nolimits_y \tau(y) \cdot \log\left(\frac{\tau(y)}{(c \gg\!= \sigma)(y)}\right) \\
&= \sum\nolimits_y \tau(y) \cdot \log\left(\frac{\tau(y)}{(c \gg\!= (c^\dagger_\sigma \gg\!= \tau))(y)} \cdot \frac{(c \gg\!= \sigma)(y)}{\tau(y)}\right) \\
&\leq \log\left(\sum\nolimits_y \frac{\tau(y) \cdot (c \gg\!= \sigma)(y)}{(c \gg\!= (c^\dagger_\sigma \gg\!= \tau))(y)}\right) \\
&\leq \log(1) \;=\; 0.
\end{aligned}$$

The first inequality is an instance of Jensen's inequality, see Lemma 2.7.3. The second one follows from:

$$\begin{aligned}
&\sum\nolimits_y \frac{\tau(y) \cdot (c \gg\!= \sigma)(y)}{(c \gg\!= (c^\dagger_\sigma \gg\!= \tau))(y)} \\
&= \sum\nolimits_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_x (c^\dagger_\sigma \gg\!= \tau)(x) \cdot c(x)(y)} \\
&= \sum\nolimits_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_{x,z} \tau(z) \cdot c^\dagger_\sigma(z)(x) \cdot (c \mathrel{=\!\!\ll} \mathbf{1}_y)(x)} \\
&\overset{(6.6)}{=} \sum\nolimits_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_{x,z} \tau(z) \cdot \frac{\sigma(x) \cdot (c \mathrel{=\!\!\ll} \mathbf{1}_z)(x)}{\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_z} \cdot (c \mathrel{=\!\!\ll} \mathbf{1}_y)(x)} \\
&= \sum\nolimits_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_z \tau(z) \cdot \frac{\sum_x \sigma(x) \cdot ((c \mathrel{=\!\!\ll} \mathbf{1}_z) \& (c \mathrel{=\!\!\ll} \mathbf{1}_y))(x)}{\sigma \models (c \mathrel{=\!\!\ll} \mathbf{1}_z) \& (c \mathrel{=\!\!\ll} \mathbf{1}_y)} \cdot \frac{\sigma \models (c \mathrel{=\!\!\ll} \mathbf{1}_z) \& (c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_z}} \\
&= \sum\nolimits_y \frac{\tau(y) \cdot (\sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_y)}{\sum_z \tau(z) \cdot (\sigma|_{c \mathrel{=\!\!\ll} \mathbf{1}_z} \models c \mathrel{=\!\!\ll} \mathbf{1}_y)} \qquad \text{by Proposition 6.1.3 (1)} \\
&\leq 1, \qquad \text{by Proposition 6.7.10.}
\end{aligned}$$

In the last line we apply Proposition 6.7.10 with test $p_i \coloneqq c \mathrel{=\!\!\ll} \mathbf{1}_{y_i}$, where $Y = \{y_1, \ldots, y_n\}$. The point predicates $\mathbf{1}_{y_i}$ form a test on Y. Predicate transformation preserves tests, see Exercise 4.3.4. As is common with daggers $c^\dagger_\sigma$, we assume that $c \gg\!= \sigma$ has full support, so that $\sigma \models p_i = \sigma \models c \mathrel{=\!\!\ll} \mathbf{1}_{y_i} = (c \gg\!= \sigma)(y_i)$ is non-zero for each i. ❑

We now turn to the proof of Proposition 6.7.10. It is relies on some basic facts from linear algebra, esp. about non-negative matrices, see [124] for background information. We shall see that for a test (p_i) the associated conditional expectations $\omega|_{p_j} \models p_i$ form a non-negative matrix C with mathematically in-

teresting properties. The proof that we present is extracted[2] from [50] and is applied only to this conditional expectation matrix C. The original proof is formulated more generally.

We recall that for a square matrix A the *spectral radius* $\rho(A)$ is the maximum of the absolute values of its eigenvalues:

$$\rho(A) \coloneqq \max\Big\{\, |\lambda| \;\Big|\; \lambda \text{ is an eigenvalue of } A \,\Big\}.$$

We shall make use the following result. The first item is known as Gelfand's formula, originally from 1941. The proof is non-trivial and is skipped here; for details, see *e.g.* [112, Appendix 10]. For convenience we include short (standard) proofs of the other two items.

Theorem 6.7.11. *Let A be a (finite) square matrix, and $\|-\|$ be a matrix norm.*

1 *The spectral radius satisfies:*

$$\rho(A) \;=\; \lim_{n\to\infty} \left\| A^n \right\|^{1/n}.$$

2 *Here we shall use the* 1*-norm* $\|A\|_1 \coloneqq \max_j \sum_i |A_{ij}|$*. It yields that* $\rho(A) = 1$ *for each (left/column-)stochastic matrix A.*

3 *Let square matrix A now be non-negative, that is, satisfy $A_{ij} \geq 0$, and let x be a positive vector, so each $x_i > 0$. If $Ax \leq r \cdot x$ with $r > 0$, then $\rho(A) \leq r$.*

Proof. As mentioned, we skip the proof of the Gelfand's formula. If A is stochastic, one gets:

$$\|A\|_1 \;=\; \max_j \textstyle\sum_i |A_{ij}| \;=\; \max_j \sum_i A_{ij} \;=\; \max_j 1 \;=\; 1.$$

Since stochastic matrices are closed under matrix multiplication, one gets $\|A^n\|_1 = 1$ for each n. Hence $\rho(A) = 1$ via Gelfand's formula.

We next show how the third item can be obtained from the first one, as in [98, Cor. 8.2.2]. By assumption, each entry x_i in the (finite) vector x is positive. Let's write x_- for the least one and x_+ for the greatest one. Then $0 < x_- \leq x_i \leq x_+$ for each i. For each n we have:

$$\begin{aligned}
\left\| A^n \right\|_1 \cdot x_- \;&=\; \max_j \textstyle\sum_i \left| \left(A^n\right)_{ij} \right| \cdot x_- \\
&\leq\; \max_j \textstyle\sum_i \left(A^n\right)_{ij} \cdot x_i \\
&=\; \max_j \left(A^n x\right)_j \\
&\leq\; \max_j \left(r^n \cdot x\right)_j \\
&\leq\; r^n \cdot x_+.
\end{aligned}$$

[2] With the help of Harald Woracek and Ana Sokolova.

Hence:

$$\left\| A^n \right\|_1^{1/n} \leq \left(r^n \cdot \frac{x_+}{x_-} \right)^{1/n} = r \cdot \left(\frac{x_+}{x_-} \right)^{1/n}.$$

Thus, by Gelfand's formula, in the first item,

$$\rho(A) = \lim_{n\to\infty} \left\| A^n \right\|^{1/n} \leq \lim_{n\to\infty} r \cdot \left(\frac{x_+}{x_-} \right)^{1/n} = r \cdot \lim_{n\to\infty} \left(\frac{x_+}{x_-} \right)^{1/n} = r \cdot 1 = r. \quad \square$$

For the remainder of this subsection the setting is as follows. Let $\omega \in \mathcal{D}(X)$ be a fixed state, with an n-test $p_1, \ldots, p_n \in \mathit{Pred}(X)$ so that $\mathbb{V}_i p_i = \mathbf{1}$, by definition (of test). We shall assume that the validities $v_i := \omega \models p_i$ are non-zero. Notice that $\sum_i v_i = 1$. We organise these validities in a vector v and in diagonal matrix V:

$$v := \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \omega \models p_1 \\ \vdots \\ \omega \models p_n \end{pmatrix} \qquad V := \begin{pmatrix} v_1 & & 0 \\ & \ddots & \\ 0 & & v_n \end{pmatrix}.$$

In addition, we use two $n \times n$ (real, non-negative) matrices B and C given by:

$$\begin{aligned} B_{ij} &:= \frac{\omega \models p_i \,\&\, p_j}{(\omega \models p_i) \cdot (\omega \models p_j)} \\ C_{ij} &:= \omega|_{p_j} \models p_i = \frac{\omega \models p_i \,\&\, p_j}{\omega \models p_j} = (\omega \models p_i) \cdot B_{ij}. \end{aligned} \tag{6.9}$$

In the last line we use Bayes' product rule from Theorem 6.1.3 (1). The next series of facts is extracted from [50].

Lemma 6.7.12. *The above matrices B and C satisfy the following properties.*

1. *The matrix B is non-negative and symmetric, and satisfies $B_{ii} \geq 1$ and $B\,v = \mathbf{1}$. Moreover, B is positive definite, so that its eigenvalues are positive reals.*
2. *As a result, the inverse B^{-1} and square root $B^{1/2}$ exist — and $B^{-1/2}$ too.*
3. *The matrix C of conditional expectations is (left/column-)stochastic and thus its spectral radius $\rho(C)$ equals 1, by Theorem 6.7.11 (2). Moreover, C satisfies $C\,v = v$ and $C = V \cdot B$.*
4. *For an $n \times n$ real matrix D, $\rho(DC) = \rho(B^{1/2} D V B^{1/2})$.*
5. *Assume now that D is a diagonal matrix with numbers $d_1, \ldots, d_n \geq 0$ on its diagonal. Then:*

$$\sum_i d_i \cdot v_i \leq \rho(DC).$$

Proof. 1 Clearly, $B_{ij} = B_{ji}$, and $B_{ii} \geq 1$ by (the byproduct in) Lemma 5.1.3. Further,

$$\begin{aligned}(B\,v)_i &= \sum_j \frac{\omega \models p_i \mathbin{\&} p_j}{(\omega \models p_i) \cdot (\omega \models p_j)} \cdot (\omega \models p_j) \\ &= \frac{\omega \models p_i \mathbin{\&} (\sum_j p_j)}{\omega \models p_i} = \frac{\omega \models p_i \mathbin{\&} \mathbf{1}}{\omega \models p_i} = \frac{\omega \models p_i}{\omega \models p_i} = 1.\end{aligned}$$

The matrix B is positive definite since for a non-zero vector $z = (z_i)$ of reals:

$$\begin{aligned}z^T B\,z &= \sum_{i,j} z_i \cdot \frac{\omega \models p_i \mathbin{\&} p_j}{(\omega \models p_i) \cdot (\omega \models p_j)} \cdot z_j \\ &= \omega \models \left(\textstyle\sum_i \frac{z_i}{v_i} \cdot p_i\right) \mathbin{\&} \left(\sum_j \frac{z_j}{v_j} \cdot p_j\right) \\ &= \omega \models q \mathbin{\&} q \qquad \text{for } q = \textstyle\sum_i \frac{z_i}{v_i} \cdot p_i \\ &> 0.\end{aligned}$$

We have a strict inequality $>$ here since $q \geq \frac{z_1}{v_1} \cdot p_1$ and $\omega \models p_1 > 0$, by assumption; in fact this holds for each p_i. Thus:

$$\omega \models q \mathbin{\&} q \geq \tfrac{z_1^2}{v_1^2} \cdot (\omega \models p_1 \mathbin{\&} p_1) \geq \tfrac{z_1^2}{v_1^2} \cdot (\omega \models p_1)^2 > 0.$$

2 The square root $B^{1/2}$ and inverse B^{-1} are obtained in the standard way via spectral decomposition $B = Q\Lambda Q^T$ where Λ is the diagonal matrix of eigenvalues $\lambda_i > 0$ and Q is an orthogonal matrix (so $Q^T = Q^{-1}$). Then: $B^{1/2} = Q\Lambda^{1/2}Q^T$ where $\Lambda^{1/2}$ has entries $\lambda_i^{1/2}$. Similarly, $B^{-1} = Q\Lambda^{-1}Q^T$, and $B^{-1/2} = Q\Lambda^{-1/2}Q^T$.

3 It is easy to see that all C's columns add up to one:

$$\textstyle\sum_i C_{ij} = \sum_i \omega|_{p_j} \models p_i = \omega|_{p_j} \models \sum_i p_i = \omega|_{p_j} \models \mathbf{1} = 1.$$

This makes C left-stochastic, so that $\rho(C) = 1$. Next:

$$\begin{aligned}(C\,v)_i &= \textstyle\sum_j (\omega|_{p_j} \models p_i) \cdot (\omega \models p_j) \\ &= \textstyle\sum_j \omega \models p_i \mathbin{\&} p_j \qquad \text{by Proposition 6.1.3 (1)} \\ &= \omega \models p_i \mathbin{\&} (\textstyle\sum_j p_j) = \omega \models p_i \mathbin{\&} \mathbf{1} = \omega \models p_i = v_i.\end{aligned}$$

Further, $(VB)_{ij} = v_i \cdot B_{ij} = C_{ij}$.

4 We show that DC and $B^{1/2}DVB^{1/2}$ have the same eigenvalues, which gives $\rho(DC) = \rho(B^{1/2}DVB^{1/2})$. First, let $DC\,z = \lambda z$. Take $z' = B^{1/2}z$ one gets:

$$B^{1/2}DVB^{1/2}z' = B^{1/2}DVB^{1/2}B^{1/2}z = B^{1/2}DCz = B^{1/2}\lambda z = \lambda B^{1/2}z = \lambda z'.$$

In the other direction, let $B^{1/2}DVB^{1/2}w = \lambda w$. Now take $w' = B^{-1/2}w$ so that:

$$DCw' = B^{-1/2}B^{1/2}DVBB^{-1/2}w = B^{-1/2}(B^{1/2}DVB^{1/2})w = B^{-1/2}\lambda w = \lambda w'.$$

5 We use the standard fact that for non-zero vectors z one has:

$$\frac{|(Az, z)|}{(z, z)} \leq \rho(A),$$

where $(-,-)$ is inner product. In particular,

$$\frac{|(B^{1/2}DVB^{1/2}z, z)|}{(z, z)} \leq \rho(B^{1/2}DVB^{1/2}) = \rho(DC).$$

We instantiate with $z = B^{1/2}v$ and use that $VBv = Cv = v$ and $Bv = \mathbf{1}$ in:

$$\begin{aligned}\rho(DC) &\geq \frac{|(B^{1/2}DVB^{1/2}B^{1/2}v, B^{1/2}v)|}{(B^{1/2}v, B^{1/2}v)} = \frac{|(DVBv, B^{1/2}B^{1/2}v)|}{(v, B^{1/2}B^{1/2}v)} \\ &= \frac{|(Dv, Bv)|}{(v, Bv)} = \frac{|(Dv, \mathbf{1})|}{(v, \mathbf{1})} = \frac{|\sum_i d_i \cdot v_i|}{\sum_i v_i} = \sum_i d_i \cdot v_i. \qquad \blacksquare\end{aligned}$$

We are now close to proving the inequality of Proposition 6.7.10 — which is the aim of this subsection. Consider a vector r of non-zero numbers $r_i \in (0, 1]$ with $\sum_i r_i = 1$. We form a diagonal matrix D with non-zero diagonal entries $d_1, \ldots, d_n$ with:

$$d_i := \frac{r_i}{(Cr)_i} = \frac{r_i}{\sum_j C_{ij} \cdot r_j}.$$

A crucial observation is that r is an eigenvector of the matrix DC, with eigenvalue 1, since:

$$(DCr)_i = \sum_j (DC)_{ij} \cdot r_j = \sum_j d_i \cdot C_{ij} \cdot r_j = \frac{r_i}{(Cr)_i} \cdot \Big(\sum_j C_{ij} \cdot r_j\Big) = r_i.$$

Theorem 6.7.11 (3) now yields $\rho(DC) = 1$.

By Lemma 6.7.12 (5) we get the required inequality in Proposition 6.7.10:

$$\sum_i \frac{r_i \cdot v_i}{(Cr)_i} = \sum_i d_i \cdot v_i \leq \rho(DC) = 1.$$

Exercises

6.7.1 Check for yourself the claimed outcomes in Example 6.7.5:

1 $\omega|_{\pi_2 \ll \mathbf{1}_e}[1, 0] = \frac{4}{5}|c\rangle + \frac{1}{5}|c^\perp\rangle$;
2 $\omega' := (\pi_1)^\dagger_\omega \gg\!= \rho = \frac{1}{10}|c, e\rangle + \frac{1}{40}|c, e^\perp\rangle + \frac{7}{40}|c^\perp, e\rangle + \frac{7}{10}|c^\perp, e^\perp\rangle$;
3 $\omega'|_{\pi_2 \ll \mathbf{1}_e}[1, 0] = \frac{4}{11}|c\rangle + \frac{7}{11}|c^\perp\rangle$.

6.7.2 This example is taken from [39], where it is attributed to Whitworth: there are three contenders A, B, C for winning a race with a prori distribution $\omega = \frac{2}{11}|A\rangle + \frac{4}{11}|B\rangle + \frac{5}{11}|C\rangle$. Surprising information comes in that A's chances have become $\frac{1}{2}$. What are the adapted chances of B and C?

1 Split up the sample space $X = \{A, B, C\}$ into a suitable two-element partition via a function $f\colon X \to 2$.

2 Use the uniform distribution $\mathit{unif} = \frac{1}{2}|1\rangle + \frac{1}{2}|0\rangle$ on 2 and show that Jeffrey's rule gives as adapted distribution:

$$f^{\dagger}_{\omega} \gg\!= \mathit{unif} \;=\; \tfrac{1}{2} \cdot \omega|_{\mathbf{1}_{\{A\}}} + \tfrac{1}{2} \cdot \omega|_{\mathbf{1}_{\{B,C\}}} \;=\; \tfrac{1}{2}|A\rangle + \tfrac{2}{9}|B\rangle + \tfrac{5}{18}|C\rangle.$$

6.7.3 The next illustration is attributed to Jeffrey, and reproduced for instance in [19, 33]. We consider three colors: green (g), blue (b) and violet (v), which are combined in a space $C = \{g, b, v\}$. These colors apply to cloths, which can additionally be sold or not, as represented by the space $S = \{s, s^{\perp}\}$. There is a prior joint distribution τ on $C \times S$, namely:

$$\omega \;=\; \tfrac{3}{25}|g, s\rangle + \tfrac{9}{50}|g, s^{\perp}\rangle + \tfrac{3}{25}|b, s\rangle + \tfrac{9}{50}|b, s^{\perp}\rangle + \tfrac{8}{25}|v, s\rangle + \tfrac{2}{25}|v, s^{\perp}\rangle.$$

A cloth is inspected by candlelight and the following likelihoods are reported per color: 70% certainty that it is green, 25% that it is blue, and 5% that it is violet.

1 Compute the two marginals $\omega_1 \coloneqq \omega[1, 0] \in \mathcal{D}(C)$ and $\omega_2 \coloneqq \omega[0, 1] \in \mathcal{D}(S)$ and show that we can write the joint state ω in two ways as:

$$\langle c,\, \mathit{id}\rangle \gg\!= \omega_2 \;=\; \omega \;=\; \langle \mathit{id},\, d\rangle \gg\!= \omega_1$$

for channels $c\colon S \nrightarrow C$ and $d\colon C \nrightarrow S$, given by:

$$\begin{cases} c(s) \;=\; \frac{3}{14}|b\rangle + \frac{3}{14}|b\rangle + \frac{8}{14}|v\rangle \\ c(s^{\perp}) \;=\; \frac{9}{22}|b\rangle + \frac{9}{22}|b\rangle + \frac{2}{11}|v\rangle \end{cases} \qquad \begin{cases} d(g) \;=\; \frac{2}{5}|s\rangle + \frac{3}{5}|s^{\perp}\rangle \\ d(b) \;=\; \frac{2}{5}|s\rangle + \frac{3}{5}|s^{\perp}\rangle \\ d(v) \;=\; \frac{4}{5}|s\rangle + \frac{1}{5}|s^{\perp}\rangle. \end{cases}$$

These channels c, d are each other's daggers, see Theorem 6.6.5 (1).

2 Capture the above inspection evidence as a predicate q on C and show that Pearl's rule gives:

$$\omega_2|_{c \gg\!\prec q} \;=\; (\omega|_{q \otimes \mathbf{1}})[0, 1] \;=\; d \gg\!= (\omega_1|_q) \;=\; \tfrac{26}{61}|s\rangle + \tfrac{35}{61}|s^{\perp}\rangle.$$

3 Describe the evidence now as a state τ on C and show that Jeffrey's rule gives:

$$d \gg\!= \tau \;=\; \big((\pi_1)^{\dagger}_{\omega} \gg\!= \tau\big)[0, 1] \;=\; \tfrac{21}{50}|s\rangle + \tfrac{29}{50}|s^{\perp}\rangle.$$

4 Check also that:

$$\begin{aligned} \langle \mathit{id}, c\rangle \gg\!= \tau \;=\; (\pi_1)^{\dagger}_{\omega} \gg\!= \tau \;=\; \tfrac{7}{25}|g, s\rangle + \tfrac{21}{50}|g, s^{\perp}\rangle) + \tfrac{1}{10}|b, s\rangle \\ + \tfrac{3}{20}|b, s^{\perp}\rangle) + \tfrac{1}{25}|v, s\rangle + \tfrac{1}{100}|v, s^{\perp}\rangle. \end{aligned}$$

The latter outcome is given in [33].

6.7.4 The following alarm example is in essence due to Pearl [133], see also [33, §3.6]; we have adapted the numbers in order to make the calculations a little bit easier. There is an 'alarm' set $A = \{a, a^\perp\}$ and a 'burglary' set $B = \{b, b^\perp\}$, with the following a priori joint distribution on $A \times B$.

$$\omega \;=\; \tfrac{1}{200}|a, b\rangle + \tfrac{7}{500}|a, b^\perp\rangle + \tfrac{1}{1000}|a^\perp, b\rangle + \tfrac{98}{100}|a^\perp, b^\perp\rangle.$$

Someone reports that the alarm went off, but with only 80% certainty because of deafness.

1 Translate the alarm information into a predicate $p\colon A \to [0, 1]$ and show that crossover updating leads to a burglary distribution:

$$\begin{aligned}\omega|_{p\otimes\mathbf{1}}[0, 1] \;&=\; \tfrac{3}{151}|b\rangle + \tfrac{148}{151}|b^\perp\rangle \\ &\approx\; 0.02|b\rangle + 0.98|b^\perp\rangle.\end{aligned}$$

2 Compute the extracted channel $c \coloneqq \pi_2 \circ (\pi_1)^\dagger_\omega\colon B \nrightarrow A$ as in Theorem 6.6.5, and express the answer in the previous item in terms of Pearl's update rule / backward inference using c.

3 Use the Bayesian inversion/dagger $d \coloneqq c^\dagger_{\omega[0,1]}\colon A \nrightarrow B$ of this channel c to calculate the outcome of Jeffrey's update rule as:

$$\begin{aligned}d \gg\!= (\tfrac{4}{5}|a\rangle + \tfrac{1}{5}|a^\perp\rangle) \;&=\; \tfrac{19639}{93195}|b\rangle + \tfrac{73556}{93195}|b^\perp\rangle \\ &\approx\; 0.21|b\rangle + 0.79|b^\perp\rangle.\end{aligned}$$

(Notice again the considerable difference in outcomes between Pearl and Jeffrey.)

6.7.5 Recall from Example 6.7.2 the Jeffrey update:

$$\sigma_1 \;\coloneqq\; c^\dagger_\sigma \gg\!= \rho \;=\; \tfrac{278}{519}|d\rangle + \tfrac{278}{519}|d^\perp\rangle.$$

Suppose we have another evidence state $\rho = \frac{3}{5}|p\rangle + \frac{2}{5}|n\rangle$. Compute:

1 $c^\dagger_{\sigma_1} \gg\!= \rho$;

2 $\sigma_2 \coloneqq c^\dagger_\sigma \gg\!= \rho$;

3 $c^\dagger_{\sigma_2} \gg\!= \tau$.

The first and third items produce different results, which proves Proposition 6.7.6 (4).

6.7.6 Let $\omega \in \mathcal{D}(X)$ be a state.

1 Consider a predicate $p\colon X \to [0,1]$ as a channel $p\colon X \nrightarrow 2$, using that $\mathcal{D}(2) \cong [0,1]$, see Exercises 4.3.5 and 7.6.1. Show that the binary version of the law of total probability, see Exercise 6.1.4, can be expressed via Jeffrey's rule as:

$$\omega = p^{\dagger}_{\omega} \gg\!= \Big((\omega \models p)|1\rangle + (\omega \models p^{\perp})|0\rangle\Big).$$

2 Let $c\colon X \nrightarrow n$ be a channel and define n predicates on X by $p_i \coloneqq c \mathrel{=\!\ll} \mathbf{1}_i$. Show that these $p_0, \ldots, p_{n-1}$ form a test:

$$\mathbb{V}_i\, p_i = \mathbf{1} \qquad \text{with} \qquad c^{\dagger}_{\omega}(i) = \omega|_{p_i}.$$

3 Conclude that the equation $c^{\dagger}_{\omega} \gg\!= (c \gg\!= \omega) = \omega$ from Proposition 6.6.4 (1) can also be understood as the (n-ary version of the) law of total probability.

6.7.7 Consider a channels $c\colon X \nrightarrow Y$ with initial state $\sigma \in \mathcal{D}(X)$ and evidence state $\tau \in \mathcal{D}(Y)$. Assume that τ is a point state $1|z\rangle$, for some $z \in Y$. Write $\omega \coloneqq \langle id, c\rangle \gg\!= \sigma \in \mathcal{D}(X \times Y)$.

1 Show that the Jeffrey-updated joint state $\omega' \coloneqq (\pi_2)^{\dagger}_{\omega} \gg\!= \tau \in \mathcal{D}(X \times Y)$ is of the form:

$$\omega' = \sigma|_{c \mathrel{=\!\ll} \mathbf{1}_z} \otimes 1|z\rangle.$$

2 Verify that the double-dagger channel $c' \coloneqq (c^{\dagger}_{\sigma})^{\dagger}_{\tau}\colon X \nrightarrow Y$ is extremely trivial, namely a 'constant-point' channel, that is, of the form $c'(x) = 1|z\rangle$ for all $x \in X$.

6.7.8 Jeffrey's rule is frequently formulated (notably in [61], to which we refer for details) and used (like in Exercise 6.7.2), in situations where the channel involved is deterministic. Consider an arbitrary function $f\colon X \to I$, giving a partition of the set X via subsets $U_i \coloneqq f^{-1}(i) = \{x \in X \mid f(x) = i\}$. Let $\omega \in \mathcal{D}(X)$ be a prior.

1 Show that applying Jeffrey's rule to a new state of affairs $\rho \in \mathcal{D}(I)$ gives as posterior:

$$f^{\dagger}_{\omega} \gg\!= \rho = \sum_{i \in I} \rho(i) \cdot \omega|_{\mathbf{1}_{U_i}} \qquad \text{satisfying} \qquad f \gg\!= (f^{\dagger}_{\omega} \gg\!= \rho) = \rho.$$

2 Prove the following minimal-distance result:

$$d(f^{\dagger}_{\omega} \gg\!= \rho, \omega) = \bigwedge \{d(\omega, \omega') \mid \omega' \in \mathcal{D}(X) \text{ with } f \gg\!= \omega' = \rho\},$$

where d is the total variation distance from Section 4.5.

6.7.9 Prove that for a general, not-deterministic channel $c\colon X \nrightarrow Y$ with prior state $\omega \in \mathcal{D}(X)$ and state $\rho \in \mathcal{D}(Y)$ that there is an inequality:

$$d(c^{\dagger}_{\omega} \gg\!= \rho, \omega) \leq \bigwedge_{\omega' \in \mathcal{D}(X)} d(\omega, \omega') + d(c \gg\!= \omega', \rho).$$

6.7.10 Show that:

$$\begin{aligned} D_{KL}(c^{\dagger}_{\sigma} \gg\!= \tau,\, \sigma) &\leq D_{KL}(\tau,\ c \gg\!= \sigma) \\ D_{KL}(\sigma,\, c^{\dagger}_{\sigma} \gg\!= \tau) &\leq D_{KL}(c \gg\!= \sigma,\, \tau). \end{aligned}$$

6.7.11 Consider the set $\{a, b, c\}$ with distribution and test of predicates:

$$\omega = \tfrac{1}{4}|a\rangle + \tfrac{1}{2}|b\rangle + \tfrac{1}{4}|c\rangle \qquad \begin{aligned} p_1 &= \tfrac{1}{2} \cdot \mathbf{1}_a + \tfrac{1}{2} \cdot \mathbf{1}_b \\ p_2 &= \tfrac{1}{2} \cdot \mathbf{1}_a + \tfrac{1}{4} \cdot \mathbf{1}_b + \tfrac{1}{2} \cdot \mathbf{1}_c \\ p_3 &= \tfrac{1}{4} \cdot \mathbf{1}_b + \tfrac{1}{2} \cdot \mathbf{1}_c. \end{aligned}$$

1 Check that the two matrices B, C defined in (6.9) are:

$$B = \begin{pmatrix} 4/3 & 8/9 & 2/3 \\ 8/9 & 10/9 & 1 \\ 2/3 & 1 & 3/2 \end{pmatrix} \qquad \text{and} \qquad C = \begin{pmatrix} 1/2 & 1/3 & 1/4 \\ 1/3 & 5/12 & 3/8 \\ 1/6 & 1/4 & 3/8 \end{pmatrix}.$$

2 Show that C has eigenvalues 1, of course, and $\frac{21+3\sqrt{17}}{144} \approx 0.2317$ and $\frac{21-3\sqrt{17}}{144} \approx 0.0599$, both below 1.

3 Check that the inequality in Proposition 6.7.10 amounts to: for positive $r_1, r_2, r_3 \in (0, 1]$ with $r_1 + r_2 + r_3 = 1$ one has:

$$\begin{aligned} 1 &\geq \frac{\frac{3}{8}r_1}{\frac{1}{2}r_1 + \frac{1}{3}r_2 + \frac{1}{4}r_3} + \frac{\frac{3}{8}r_2}{\frac{1}{3}r_1 + \frac{5}{12}r_2 + \frac{3}{8}r_3} + \frac{\frac{1}{4}r_3}{\frac{1}{6}r_1 + \frac{1}{4}r_2 + \frac{3}{8}r_3} \\ &= \frac{9r_1}{12r_1 + 8r_2 + 6r_3} + \frac{9r_2}{8r_1 + 10r_2 + 9r_3} + \frac{6r_3}{4r_1 + 6r_2 + 9r_3}. \end{aligned}$$

Try to prove this inequality via analytical means.

6.8 Frequentist and Bayesian discrete probability

Here, at the end of Chapter 6, we have seen the basic concepts in (our treatment of) discrete probability theory, namely: distributions (states), predicates, validity, updating. We shall further explore these notions in subsequent chapters **??** – **??**, but at this point we like to sit back and reflect on what we have seen so far, from a more abstract perspective.

We start with a fundamental result, for which we recall that the set $\mathit{Pred}(X) = [0, 1]^X$ of fuzzy predicates on a set X has the structure of an effect module: there

is truth and falsity $\mathbf{1}, \mathbf{0}$, orthocomplement $p^\perp$, partial sum $p \owedge q$, and scalar multiplication $r \cdot p$, see Subsection 4.2.3.

The next result is a Riesz-style representation result, representing distributions on X via a double dual $[0,1]^{[0,1]^X}$.

Theorem 6.8.1. *Let X be a finite set. There is then a 'representation' isomorphism,*

$$\mathcal{D}(X) \xrightarrow[\cong]{\mathcal{V}} \mathbf{EMod}\Big(\mathit{Pred}(X),\, [0,1]\Big) \tag{6.10}$$

given by validity: $\mathcal{V}(\omega)(p) \coloneqq \omega \models p$.

This result uses the notation $\mathbf{EMod}(\mathit{Pred}(X), [0,1])$ for the 'hom' set of homomorphisms of effect modules $\mathit{Pred}(X) \to [0,1]$.

Proof. We have already seen that $\mathcal{V}(\omega) \colon \mathit{Pred}(X) \to [0,1]$ preserves the effect module structure, see Lemma 4.2.6. In order to show that it is an isomorphism, let $h \colon \mathit{Pred}(X) \to [0,1]$ be a homomorphism of effect modules. It gives rise to a distribution:

$$\mathcal{V}^{-1}(h) \coloneqq \sum_{x \in X} h(\mathbf{1}_x) \big| x \big\rangle \in \mathcal{D}(X).$$

These $\mathcal{V}$ and $\mathcal{V}^{-1}$ are each other's inverses:

$$\begin{aligned}
(\mathcal{V}^{-1} \circ \mathcal{V})(\omega) &= \sum_{x \in X} \mathcal{V}(\omega)(\mathbf{1}_x) \big| x \big\rangle = \sum_{x \in X} (\omega \models \mathbf{1}_x) \big| x \big\rangle \\
&= \sum_{x \in X} \omega(x) \big| x \big\rangle = \omega \\
(\mathcal{V} \circ \mathcal{V}^{-1})(h)(p) &= \left(\sum_{x \in X} h(\mathbf{1}_x) \big| x \big\rangle \right) \models p = \sum_{x \in X} h(\mathbf{1}_x) \cdot p(x) \\
&= h\left(\bigowedge_{x \in X} p(x) \cdot \mathbf{1}_x \right) = h(p).
\end{aligned}$$

In the last line we use that h is a homomorphism of effect modules and that the predicate p has a normal form $\bigowedge_x p(x) \cdot \mathbf{1}_x$, see Lemma 4.2.3 (2). ❑

There are two leading interpretations of probability, namely a frequentist interpretation and a Bayesian interpretation. The frequentist approach treats probability distributions as records of probabilities of occurrences, obtained from long-term accumulations, *e.g.* via frequentist learning. One can associate this view with the set $\mathcal{D}(X)$ of distributions on the left-hand-side of the representation isomorphism (6.10). The right-hand-side fits the Bayesian view, which focuses on assigning probabilities to belief functions (predicates), see [38], in this situation via special functions $\mathit{Pred}(X) \to [0,1]$ that preserve

the effect module structure. The representation isomorphism demonstrates that the frequentist and Bayesian interpretations of probability are thightly connected, at least in finite discrete probability. We shall see a similar isomorphism below, in Theorem 6.8.4, for infinite discrete probability. The situation for continuous probability is more subtle and will be described in **??**.

But first we look at how operations on distributions, like marginalisation, tensor product, and updating work accross the above representation isomorphism (6.10). The next result establishes a close connection between operations on distributions and logical operations, like between marginalisation and weakening. In the case of updating (conditioning) of a distribution, we see that the corresponding logical formulation is the common formulation of conditonal probability as a fraction of probability assignments. Not suprisingly, the connection relies on Bayes' rule.

Proposition 6.8.2. *Fix finite sets X, Y.*

1 *Let $h\colon \mathsf{Pred}(X \times Y) \to [0,1]$ be a map of effect modules. Define first and second marginals $h[1,0]\colon \mathsf{Pred}(X) \to [0,1]$ and $h[0,1]\colon \mathsf{Pred}(Y) \to [0,1]$ of h as:*

$$h[1,0](p) \coloneqq h(p \otimes \mathbf{1}) \qquad \text{and} \qquad h[0,1](q) \coloneqq h(\mathbf{1} \otimes q).$$

Then: $\mathcal{V}^{-1}(h[1,0]) = \mathcal{V}^{-1}(h)[1,0]$ and $\mathcal{V}^{-1}(h[0,1]) = \mathcal{V}^{-1}(h)[0,1]$.

2 *Let $h\colon \mathsf{Pred}(X) \to [0,1]$ and $k\colon \mathsf{Pred}(Y) \to [0,1]$ be maps of effect modules. We define their tensor product $h \otimes k\colon \mathsf{Pred}(X \times Y) \to [0,1]$ as:*

$$\begin{aligned}(h \otimes k)(r) &\coloneqq h\Big(x \mapsto k(r(x,-))\Big) \\ &= k\Big(y \mapsto h(r(-,y))\Big).\end{aligned}$$

Then: $\mathcal{V}^{-1}(h \otimes k) = \mathcal{V}^{-1}(h) \otimes \mathcal{V}^{-1}(k)$.

3 *Let $h\colon \mathsf{Pred}(X) \to [0,1]$ be a map of effect modules and let $p \in \mathsf{Pred}(X)$ be a predicate with $h(p) \neq 0$. We define an update $h|_p\colon \mathsf{Pred}(X) \to [0,1]$ as:*

$$h|_p(q) \coloneqq \frac{h(p \mathbin{\&} q)}{h(p)}.$$

Then: $\mathcal{V}^{-1}(h|_p) = \mathcal{V}^{-1}(h)\big|_p$.

Proof. 1 We consider the first marginal only. It is easy to see that the map $h[1,0]\colon \mathsf{Pred}(X) \to [0,1]$, as defined above, preserves the effect module

structure. We get an equality of distributions since for $x \in X$,

$$\begin{aligned}\mathcal{V}^{-1}(h[1,0])(x) \;=\; h[1,0](\mathbf{1}_x) \;&=\; h(\mathbf{1}_x \otimes \mathbf{1}) \\ &= h\left(\mathbf{1}_x \otimes \bigotimes_{y\in Y} \mathbf{1}_y\right) \\ &= \sum_{y\in Y} h(\mathbf{1}_x \otimes \mathbf{1}_y) \\ &= \sum_{y\in Y} h(\mathbf{1}_{(x,y)}) \\ &= \sum_{y\in Y} \mathcal{V}^{-1}(h)(x,y) \;=\; \mathcal{V}^{-1}(h)[1,0](x).\end{aligned}$$

2 For elements $u \in X$ and $v \in Y$ we have:

$$\begin{aligned}\mathcal{V}^{-1}(h \otimes k)(u,v) \;=\; (h \otimes k)(\mathbf{1}_{(u,v)}) \;&=\; h(x \mapsto k(\mathbf{1}_{(u,v)}(x,-))) \\ &= h(x \mapsto k(\mathbf{1}_u(x) \cdot \mathbf{1}_v)) \\ &= h(x \mapsto \mathbf{1}_u(x) \cdot k(\mathbf{1}_v)) \\ &= h(k(\mathbf{1}_v) \cdot \mathbf{1}_u) \\ &= k(\mathbf{1}_v) \cdot h(\mathbf{1}_u) \\ &= \mathcal{V}^{-1}(h)(u) \cdot \mathcal{V}^{-1}(k)(v) \\ &= \Big(\mathcal{V}^{-1}(h) \otimes \mathcal{V}^{-1}(k)\Big)(u,v).\end{aligned}$$

The same can be shown for the other formulation of $h \otimes k$ in item (2).

3 Again, $h|_p \colon \mathit{Pred}(X) \to [0,1]$ is a map of effect modules. Next, for $x \in X$,

$$\begin{aligned}\mathcal{V}^{-1}(h|_p)(x) \;=\; h|_p(\mathbf{1}_x) \;&=\; \frac{h(p \otimes \mathbf{1}_x)}{h(p)} \\ &= \frac{\mathcal{V}^{-1}(h) \models p \otimes \mathbf{1}_x}{\mathcal{V}^{-1}(h) \models p} \\ &= \mathcal{V}^{-1}(h)\Big|_p \models \mathbf{1}_x \quad \text{by Bayes, see Theorem 6.1.3 (2)} \\ &= \mathcal{V}^{-1}(h)\Big|_p(x). \qquad \text{❑}\end{aligned}$$

In order to extend this result to discrete distributions in $\mathcal{D}_\infty(X)$ with possibly infinite support we need to add ω-joins to effect algebras/modules.

Definition 6.8.3. 1 A sequence (or chain) of elements $a_n \in A$, for $n \in \mathbb{N}$, in a poset A is called ascending if $a_n \leq a_{n+1}$ for each n. One says that the poset A is ω-complete if each ascending chain (a_n) has a join $\bigvee_n a_n \in A$.

2 A monotone function $f \colon A \to B$ between ω-complete posets A, B is called ω-continuous if it preserves joins of ascending chains, that is, if $f(\bigvee_n a_n) = \bigvee_n f(a_n)$, for each ascending chain (a_n) in A.

3 An ω-effect algebra is an effect algebra which is ω-complete as a poset, using the order described in Exercise 4.2.14. We shall write ω-**EA** for the category with ω-effect algebras as objects, and with ω-continuous effect algebra maps as morphisms.

4 Similarly, an ω-effect module is an ω-complete effect module. The category ω-**EMod** contains such ω-effect modules, with ω-continous effect module maps between them.

For each set X, the powerset $\mathcal{P}(X)$ is an effect algebra with arbitrary joins, so in particular ω-joins. The unit interval $[0, 1]$ is an ω-effect module, and more generally, each set of predicates $Pred(X) = [0, 1]^X$ is an effect module, via pointwise joins. Later on, in continuous probability, we shall see measurable spaces whose sets of measurable subsets form examples of ω-effect algebras.

We recall from Exercise 4.2.15 that the sum operation $\owedge$ of an ω-effect algebra is continuous in both arguments separately. Explicitly: $\bigvee_n x \owedge y_n \leq x \owedge \bigvee_n y_n$. Further, if there are joins $\bigvee$, there are also meets $\bigwedge$ via orthosupplement.

Theorem 6.8.4. *For each countable set X there is a representation isomorphism,*

$$\mathcal{D}_\infty(X) \xrightarrow[\cong]{\mathcal{V}} \omega\text{-}\mathbf{EMod}\Big(Pred(X),\, [0,1]\Big) \tag{6.11}$$

also given by validity: $\mathcal{V}(\omega)(p) \coloneqq \omega \models p$.

Proof. Much of this works as in the proof of Theorem 6.8.1, except that we have to prove that $\mathcal{V}(\omega)\colon Pred(X) \to [0, 1]$ is ω-continuous, now that we have $\omega \in \mathcal{D}_\infty(X)$. So let (p_n) be an ascending chain of predicates, with pointwise join $p = \bigvee_n p_n$. Then:

$$\begin{aligned}
\mathcal{V}(\omega)\Big(\bigvee\nolimits_n p_n\Big) \;=\; \omega \models \bigvee\nolimits_n p_n \;&=\; \sum_{x\in X} \omega(x) \cdot \Big(\bigvee\nolimits_n p_n\Big)(x) \\
&= \sum_{x\in X} \omega(x) \cdot \Big(\bigvee\nolimits_n p_n(x)\Big) \\
&= \sum_{x\in X} \bigvee\nolimits_n \omega(x) \cdot p_n(x) \\
&\overset{(*)}{=} \bigvee\nolimits_n \sum_{x\in X} \omega(x) \cdot p_n(x) \;=\; \bigvee\nolimits_n \mathcal{V}(\omega)(p_n).
\end{aligned}$$

The direction $(\geq)$ of the marked equation $\overset{(*)}{=}$ holds by monotonicity. For $(\leq)$ we reason as follows. Since X is countable, we can write it as $X = \{x_1, x_2, \ldots\}$. For

each $N \in \mathbb{N}$ we can use that finite sums preserve ω-joins in:

$$\sum_{k\leq N} \bigvee_n \omega(x_k) \cdot p_n(x_k) \;=\; \bigvee_n \sum_{k\leq N} \omega(x_k) \cdot p_n(x_k) \;\leq\; \bigvee_n \sum_{x\in X} \omega(x) \cdot p_n(x).$$

Hence:

$$\sum_{x\in X} \bigvee_n \omega(x) \cdot p_n(x) \;=\; \lim_{N\to\infty} \sum_{k\leq N} \bigvee_n \omega(x_k) \cdot p_n(x_k) \;\leq\; \bigvee_n \sum_{x\in X} \omega(x) \cdot p_n(x).$$

The inverse of $\mathcal{V}$ in (6.11) is defined as countable formal sum, using that X is countable: for a homomorphism of ω-effect modules $h\colon \mathit{Pred}(X) \to [0,1]$ we take:

$$\mathcal{V}^{-1}(h) \;\coloneqq\; \sum_{x\in X} h(\mathbf{1}_x)\,\big|x\rangle.$$

In order to see that the probabilities $h(\mathbf{1}_x)$ add up to one, we write $X = \{x_1, x_2, \ldots\}$. We express the sum over these x_i via a join of an ascending chain:

$$\begin{aligned}\sum_{x\in X} h(\mathbf{1}_x) \;=\; \bigvee_n \sum_{k\leq n} h(\mathbf{1}_{x_k}) \;&=\; \bigvee_n h\left(\bigvee_{k\leq n} \mathbf{1}_{x_k}\right) \\ &=\; h\Big(\bigvee_n \mathbf{1}_{\{x_1,\ldots,x_n\}}\Big) \;=\; h(\mathbf{1}) \;=\; 1.\end{aligned}$$

The argument that $\mathcal{V}$ and $\mathcal{V}^{-1}$ are each other's inverses works essentially in the same way as in the proof of Theorem 6.8.1. ❑

Exercises

6.8.1 Consider the representation isomorphism in Theorem 6.8.1.

1 Show that a uniform distribution corresponds to the mapping that sends a predicate to its average value.

2 Show that for finite sets X, Y and maps $h \in \mathbf{EMod}(\mathit{Pred}(X), [0,1])$ and $k \in \mathbf{EMod}(\mathit{Pred}(Y), [0,1])$ one has:

$$(h \otimes k)[1, 0] \;=\; h.$$

6.8.2 Recall from Theorem 4.2.5 that $\mathit{Pred}(X)$ is the free effect module on the effect algebra $\mathcal{P}(X)$, for a finite set X.

1 Deduce from this fact that there is an isomorphism:

$$\mathbf{EA}\Big(\mathcal{P}(X),\; [0,1]\Big) \;\cong\; \mathbf{EMod}\Big(\mathit{Pred}(X),\; [0,1]\Big).$$

Describe this isomorphism in detail.

2 Conclude that an alternative version of the representation isomorphism (6.10) is:

$$\mathcal{D}(X) \xrightarrow[\cong]{} \mathbf{EA}\Big(\mathcal{P}(X),\ [0,1]\Big)$$

6.8.3 Recall the Poisson distribution $pois[\lambda] = \sum_{k\in\mathbb{N}} e^{-\lambda} \cdot \frac{\lambda^k}{k!} |k\rangle \in \mathcal{D}_\infty(\mathbb{N})$ from (2.7). Describe the corresponding homomorphism of ω-effect modules $Pred(\mathbb{N}) \to [0,1]$ via (6.11).

6.8.4 Reformulate Proposition 4.3.3 as naturality for the representation isomorphism $\mathcal{V}$ in (6.10) with respect to channels.

7

Directed Graphical Models

In the previous chapters we have seen several examples of graphs in probabilistic reasoning, see for instance Figures 2.3, 2.4 or 6.6. At this stage a more systematic, principled approach will be developed in terms of so-called *string diagrams*. The latter are directed graphs that are built up inductively from boxes (as nodes) with wires (as edges) between them. These boxes can be interpreted as probabilistic channels. The wires are typed, where a type is a finite set that is associated with a wire. These string diagrams are convenient for depicting arrangements of channels and states, and for reasoning about them.

On a theoretical level we make a clear distinction between syntax and semantics, where graphs are introduced as syntactic objects, defined inductively, like terms in some formal language based on a signature or grammar. Graphs can be given a semantics via an interpretion as suitable collections of channels. In practice, the distinction between syntax and semantics is not so sharp: we shall frequently use graphs in interpreted form, essentially by describing a structured collection of channels as interpretation of a graph. This is like what happens in (mathematical) practice with algebraic notions like group or monoid: one can precisely formulate a syntax for monoids with operation symbols that are interpreted as actual functions. But one can also use monoids as sets with specific functions. The latter form is often most convenient. But is important to be aware of the difference between syntax and semantics.

String diagrams are similar to the graphs used for Bayesian networks. A notable difference is that they have explicit operations for copying $\curlyvee$ and discarding $\top$. This makes them more expressive (and more useful). String diagrams are the language of choice to reason about a large class of mathematical models, called symmetric monoidal categories, see [149], including categories of channels (of powerset, multiset or distribution type). Once we have seen string diagrams, we will describe Bayesian networks only in string diagrammatic form. The most notable difference is that we then write copying as an explicit

operation, as we have already started doing in Section 2.6. The language of string diagrams also makes it easier to turn a Bayesian network into a joint distribution (state). String diagrams turn out to be useful not only for representation (of joint distributions) but also for reasoning (using both forward and backward inference). We shall use string diagrams also for other probabilistic models, like Markov chains and hidden Markov models in Section 7.4.

One of the main themes in this chapter is how to move back and forth between a joint state/distribution, say on A, B, and a channel $A \nrightarrow B$. The process of extracting a channel form a joint state is called *disintegration*. It is the process of turning a joint distribution $P(a, b)$ into a conditional distribution $P(b \mid a)$. It gives direction to a probabilistic relationship. It is this disintegration technique that allows us to represent a joint state, on a product of (multiple) sets, via a graph structure with multiple channels, like in Bayesian networks, by performing disintegration multiple times. More explicitly, it is via disintegration that the conditional probability tables of a Bayesian network — which, recall, correspond to channels — can be obtained from a joint distribution. The latter can be presented as a (possibly large) empirical distribution, analogously to the (small) medicine - blood pressure example in Subsections 1.4.1 and 1.4.3.

Via the correspondence between joint states and channels one can invert channels. We have already had a first look at this Bayesian inversion — as dagger of channel — in Section 6.6; we used it in Jeffrey's update rule in Section 6.7. In this chapter we take a more systematic look at inversion. It is now obtained by first turning a channel $A \nrightarrow B$ into a joint state on A, B, then into a joint state on B, A by swapping, and then again into a channel $B \nrightarrow A$. It amounts to turning a conditional probability $P(b \mid a)$ into $P(a \mid b)$, essentially by Bayes' rule. As we have seen, this channel inversion is a fundamental operation in Bayesian probability, which bears resemblance adjoint transpose $(-)^\dagger$ of complex matrices in quantum theory. The previous chapter showed the role of inverted channels in probabilistic reasoning. The current chapter illustrates their role in machine learning, esp. in naive Bayesian classification and in learning a decision tree, see Section 7.7.

For readers interested in the more categorical aspects of Bayesian inversion, Section 7.8 elaborates how a suitable category of channels has a well-behaved dagger functor. This category of channels is shown to be equivalent to a category of couplings. It mimics the categorical structure underlying the relationship between functions $X \to \mathcal{P}(Y)$ and relations on $X \times Y$. This categorically-oriented section can be seen as an intermezzo that is not necessary for the remainder; it could be skipped.

Sections 7.9 and 7.10 return to (directed) graphical modelling. Section 7.9

uses string diagrams to express intrinsic graphical structure, called *shape*, in joint probability distributions. This is commonly expressed in terms of independence and d-separation, but we prefer to stick at this stage to a purely string-diagrammatic account. String diagrams are also used in Section 7.10 to give a precise account of Bayesian inference, including a (high-level) channel-based inference algorithm.

7.1 String diagrams

In Chapter 1 we have seen how (probabilistic) channels can be composed sequentially, via $\circ$, and in parallel, via $\otimes$. These operations $\circ$ and $\otimes$ satisfy certain algebraic laws, such as associativity of $\circ$. Parallel composition $\otimes$ is also associative, in a suitable sense, but the equations involved are non-trivial to work with. These equations are formalised via the notion of *symmetric monoidal* category that axiomatises the combination of sequential and parallel composition.

Instead of working with such somewhat complicated categories we shall work with *string diagrams*. These diagrams form a graphical language that offers a more convenient and intuitive way of reasoning with sequential and parallel composition. In particular, these diagrams allow obvious forms of topological stretching and shifting that are less complicated, and more intuitive than equations.

The main reference for string diagrams is [149]. Here we present a more focused account, concentrating on what we need in a probabilistic setting. In this first section we describe string diagrams with directed edges. In the next section we simply write undirected edges | instead of directed edges $\uparrow$, with the convention that information always flows upward. However, we have to keep in mind that we do this for convenience only, and that the graphs involved are actually directed.

String diagrams look a bit like Bayesian networks, but differ in subtle ways, esp. wrt. copying, discarding and joint states, see Subsection 7.1.1. The shift that we are making from Bayesian networks to string diagrams was already prepared in Subsection 2.6.1.

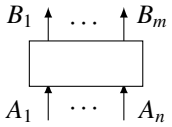
Definition 7.1.1. The collection *Type* of types is the smallest set with:

1 $1 \in \mathit{Type}$, where 1 represents the trivial singleton type;
2 $F \in \mathit{Type}$ for any finite set F;
3 if $A, B \in \mathit{Type}$, then also $A \times B \in \mathit{Type}$.

In fact, the first item can easily be seen as a special case of the second one.

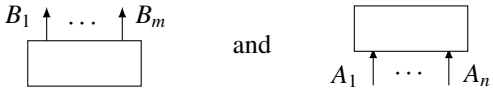
However, we like to emphasise that there is a distinguished type that we write 1. More generally, we write, as before, $n = \{0, 1, \ldots, n-1\}$, so that $n \in \mathit{Type}$.

The nodes of a string diagram are 'boxes' with input and output wires:



(7.1)

The symbols $A_1, \ldots, A_n$ are the types of the input wires, and $B_1, \ldots, B_m$ are the types of the output wires. We allow the cases $n = 0$ (no input wires) and $m = 0$ (no output wires), which are written as:



A *diagram signature*, or simply a *signature*, is a collection of boxes as in (7.1). The interpretation of such a box is a probabilistic channel $A_1 \times \cdots \times A_n \nrightarrow B_1 \times \cdots \times B_m$. In particular, the interpretation of a box with no input channels is a distribution/state. In principle, we like to make a careful distinction between syntax (boxes in a diagram) and semantics (their interpretation as channels).

Informally, a string diagram is a diagram that is built-up inductively, by connecting wires of the same type between several boxes, see for instance the two diagrams (7.3) below. The connections may include copying and swapping, see below for details. A string diagram involves the requirement that it contains no cycles: it should be *acyclic*. We write string diagrams with arrows flow going upwards, suggesting that there is a flow from bottom to top. This direction is purely a convention. Elsewhere in the literature the flow may be from top to bottom, of from left to right.

More formally, given a signature Σ of boxes of the form (7.1), the set $SD(\Sigma)$ of *string diagrams* over Σ is defined as the smallest set satisfying the items below. At the same time we define two 'domain' and 'codomain' functions from string diagrams to lists of types, as in:

$$SD(\Sigma) \underset{cod}{\overset{dom}{\rightrightarrows}} \mathcal{L}(\mathit{Type})$$

Recall that $\mathcal{L}$ is the list functor.

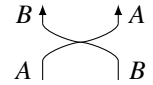
(1) $\Sigma \subseteq SD(\Sigma)$, that is, every box in $f \in \Sigma$ in the signature as in (7.1) is a string diagram in itself, with $dom(f) = [A_1, \ldots, A_n]$ and $cod(f) = [B_1, \ldots, B_m]$.

(2) For all $A, B \in \mathit{Type}$, the following diagrams are in $SD(\Sigma)$.

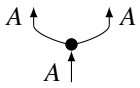
(a) The *identity* diagram:

$$id_A \;=\; \text{} \qquad \text{with} \qquad \begin{cases} dom(id_A) = [A] \\ cod(id_A) = [A]. \end{cases}$$

(b) The *swap* diagram:

$$swap_{A,B} \;=\; \text{} \qquad \text{with} \qquad \begin{cases} dom(swap_{A,B}) = [A, B] \\ cod(swap_{A,B}) = [B, A]. \end{cases}$$

(c) The *copy* diagram:

$$copy_A \;=\; \text{} \qquad \text{with} \qquad \begin{cases} dom(copy_A) = [A] \\ cod(copy_A) = [A, A]. \end{cases}$$

(d) The *discard* diagram:

$$discard_A \;=\; \text{} \qquad \text{with} \qquad \begin{cases} dom(discard_A) = [A] \\ cod(discard_A) = []. \end{cases}$$

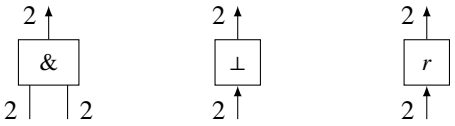
(e) The *uniform* state diagram:

$$unif_A \;=\; \text{} \qquad \text{with} \qquad \begin{cases} dom(unif_A) = [] \\ cod(unif_A) = [A]. \end{cases}$$

(f) For each type A and element $a \in A$ a *point* state diagram:

$$point_A(a) \;=\; \text{} \qquad \text{with} \qquad \begin{cases} dom(point_A(a)) = [] \\ cod(point_A(a)) = [A]. \end{cases}$$

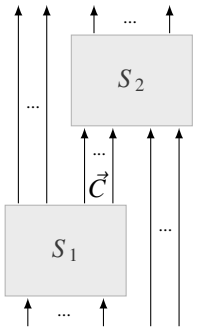
(g) Boxes for the logical operations conjunction, orthosupplement and scalar multiplication, of the form:



where $r \in [0, 1]$. These string diagrams have obvious domains and codomains. We recall that predicates on A can be identified with channels $A \nrightarrow 2$, see Exercise 4.3.5. Hence they can be included in string diagrams.

(3) If $S_1, S_2 \in SD(\Sigma)$ are string diagrams, then the *parallel composition* (juxtaposition) $S_1\, S_2$ is also string diagram, with $dom(S_1\, S_2) = dom(S_1) +\!\!+ dom(S_2)$ and $cod(S_1\, S_2) = cod(S_1) +\!\!+ cod(S_2)$.

(4) If $S_1, S_2 \in SD(\Sigma)$ and $cod(S_1)$ ends with $\vec{C} = [C_1, \ldots, C_k]$ whereas $dom(S_2)$ starts with $\vec{C}$, then there is a *sequential composition* string diagram T of the form:

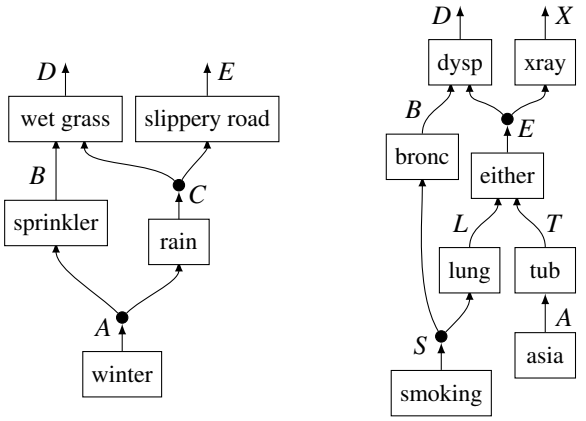
$T =$ 

with $\begin{cases} dom(T) \\ \quad = dom(S_1) +\!\!+ (dom(S_2) - \vec{C}) \\ cod(T) \\ \quad = (cod(S_1) - \vec{C}) +\!\!+ cod(S_2). \end{cases}$ (7.2)

From this basic form, different forms of composition can be obtained by reordering wires via swapping.

In this definition of string diagrams all wires point upwards. This excludes that string diagrams contain cycles: they are acyclic by construction. One could add more basic constructions to item (2), such as convex combination, see Exercise 7.2.3, but the above constructions suffices for now.

Below we redescribe the 'wetness' and 'Asia' Bayesian networks, from Figures 2.4 and 6.6, as string diagrams.



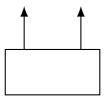
(7.3)

What are the signatures for these string diagrams?

7.1.1 String diagrams versus Bayesian networks

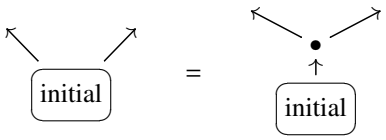
One can ask: why are these strings diagrams any better than the diagrams that we have seen so far for Bayesian networks? An important point is that string

diagrams make it possible to write a joint state, on an n-ary product type. For instance, for $n = 2$ we can have a write a joint state as:



(7.4)

In Bayesian networks joint states cannot be expressed. It is possible to write an initial node with multiple outgoing wires, but as we have seen in Section 2.6, this should be interpreted as copy of a single outgoing wire, coming out of a non-joint state:



In string diagrams one can additionally express marginals via discarding $\top$: if a joint state ω is the interpretation of the above string diagram (7.4), then the diagram on the left below is interpreted as the marginal $\omega[1, 0]$, whereas the diagram on the right is interpreted as $\omega[0, 1]$.



We shall use this marginal notation with masks also for boxes in general — and not just for states, without incoming wires. This is in line with Definition 2.5.2 where marginalisation is defined both for states and for channels.

An additional advantage of string diagrams is that they can be used for equational reasoning. This will be explained in the next section. First we give meaning to string diagrams.

7.1.2 Interpreting string diagrams as channels

String diagrams will be used as representations of (probabilistic) channels. We thus consider string diagrams as a special, graphical term calculus (syntax). These diagrams are formal constructions, which are given mathematical meaning via their interpretation as channels.

This interpretation of string diagrams can be defined in a compositional way, following the structure of a diagram, using sequential $\circ$ and parallel $\otimes$ composition of channels. An interpretation of a string diagram $S \in SD(\Sigma)$ is a channel that is written as $[\![S]\!]$. It forms a channel from the product of the types in the domain $dom(S)$ of the string diagram to the product of types in the codomain

$cod(S)$. The product over the empty list is the singleton element $1 = \{0\}$. Such an interpretation $[\![S]\!]$ is parametrised by an interpretation of the boxes in the signature Σ as channels.

Given such an interpretation of the boxes in Σ, the items below describe this interpretation function $[\![-]\!]$ inductively, acting on the whole of $SD(\Sigma)$, by precisely following the above four items in the inductive definition of string diagrams.

(1) The interpretation of a box (7.1) in Σ is assumed. It is some channel of type $A_1 \times \cdots \times A_n \to\!\!\!\!\bullet\; B_1 \times \cdots \times B_m$.
(2) We use the following interpretation of primitive boxes.
   (a) The interpretation of the identity wire of type A is the identity/unit channel $id : A \to\!\!\!\!\bullet\; A$, given by point states: $id(a) = 1|a\rangle$.
   (b) The swap string diagram is interpreted as the (switched) pair of projections $\langle \pi_2, \pi_1 \rangle : A \times B \to\!\!\!\!\bullet\; B \times A$. It sends (a, b) to $1|b, a\rangle$.
   (c) The interpretation $[\![copy_A]\!]$ of the copy string diagram is the copy channel $\Delta : A \to\!\!\!\!\bullet\; A \times A$ with $\Delta(a) = 1|a, a\rangle$.
   (d) The discard channel $\top$ of type A is interpreted as the unique channel $! : A \to\!\!\!\!\bullet\; 1$. Using that $\mathcal{D}(1) \cong 1$, this function sends every element $a \in A$ to the sole element $1|0\rangle$ of $\mathcal{D}(1)$ for $1 = \{0\}$.
   (e) The uniform string diagram $\perp$ of type A is interpreted as the uniform distribution/state $unif_A \in \mathcal{D}(A)$, considered as channel $1 \to\!\!\!\!\bullet\; A$. Recall, if A has n-elements, then $unif_A = \sum_{a \in A} \frac{1}{n}|a\rangle$.
   (f) The point string diagram $point_A(a)$, for $a \in A$ is interpreted as the point state $1|a\rangle \in \mathcal{D}(A)$.
   (g) The 'logical' channels for conjunction, orthosupplement, and scalar multiplication are interpreted by the corresponding channels $conj$, $orth$ and $scal(r)$ from Exercise 4.3.5.
(3) The tensor product $\otimes$ of channels is used to interpret juxtaposition of string diagrams $[\![S_1\, S_2]\!] = [\![S_1]\!] \otimes [\![S_2]\!]$.
(4) If we have string diagrams S_1, S_2 in a composition T as in (7.2), then:

$$[\![T]\!] \;=\; (id \otimes [\![S_2]\!]) \circ ([\![S_1]\!] \otimes id),$$

   where the identity channels have the appropriate product types.

At this stage we can say more precisely what a Bayesian network is, within the setting of string diagrams. This string diagrammatic perspective started in [48].

Definition 7.1.2. A *Bayesian network* is given by:

1 a string diagram $G \in SD(\Sigma)$ over a finite signature Σ, with $dom(G) = []$;
2 an interpretation of each box in Σ as a channel.

The string diagram G in the first item is commonly referred to as the (underlying) graph of the Bayesian network. The channels that are used as interpretations are the conditional probability tables of the network.

Given the definition of the boxes in (the signature of) a Bayesian network, the whole network can be interpreted as a state/distribution $[\![G]\!]$. This is in general not the same thing as the joint distribution associated with the network, see Definition 7.3.1 (3) later on. In addition, for each node A in G, we can look at the subgraph G_A of cumulative parents of A. The interpretation $[\![G_A]\!]$ is then also a state, namely the one that one obtains via state transformation. This has been described as 'prediction' in Section 2.6.

The above definition of Bayesian network is more general than usual, for instance because it allows the graph to contain joint states, like in (7.4), or discarding $\overline{\top}$, see also the discussion in [82].

Exercises

7.1.1 For a state $\omega \in \mathcal{D}(X_1 \times X_2 \times X_3 \times X_4 \times X_5)$, describe its marginalisation $\omega[1, 0, 1, 0, 0]$ as string diagram.

7.1.2 Check in detail how the two string diagrams in (7.3) are obtained by following the construction rules for string diagrams.

7.1.3 Consider the boxes in the wetness string diagram in (7.3) as a signature. An interpretation of the elements in this signature is described in Section 2.6 as states and channels *wi*, *sp etc.* Check that this interpretation extends to the whole wetness string diagram in (7.3) as a channel $1 \nrightarrow D \times E$ given by:

$$(wg \otimes sr) \circ (id \otimes \Delta) \circ (sp \otimes ra) \circ \Delta \circ wi.$$

7.1.4 Extend the interpretation from Section 6.5 of the signature of the Asia string diagram in (7.3) to a similar interpretation of the whole Asia string diagram.

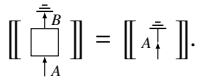
7.1.5 For $A = \{a, b, c\}$ check that:

$$[\![\; \overset{A}{\bullet} \;]\!] = \tfrac{1}{3}|a, a\rangle + \tfrac{1}{3}|b, b\rangle + \tfrac{1}{3}|c, c\rangle.$$

7.1.6 Check that there are equalities of interpretations:

$$[\![\perp^A \perp^B]\!] = [\![\perp^{A \times B}]\!] \qquad [\![\overline{\top}_A \overline{\top}_B]\!] = [\![\overline{\top}_{A \times B}]\!]$$

7.1.7 Verify that for each box, interpreted as a channel, one has:

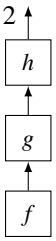


7.1.8 Recall Exercise 4.3.6 and check that the interpretation of



is $[\![f]\!] \models [\![g]\!]$.

Similarly, check that the interpretation of

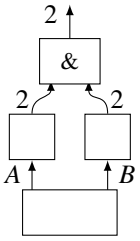


is

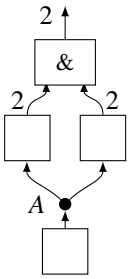
$$[\![g]\!] \gg\!= [\![f]\!] \models [\![h]\!] = [\![f]\!] \models [\![g]\!] =\!\ll [\![h]\!],$$

see Proposition 4.3.3.

7.1.9 Notice that the difference between $\omega \models p_1 \otimes p_2$ and $\sigma \models q_1 \,\&\, q_2$ can be clarified graphically as:



versus



See also Exercise 4.3.7.

7.2 Equations for string diagrams

This section introduces equations $S_1 = S_2$ between string diagrams $S_1, S_2 \in SD(\Sigma)$ over the same signature Σ. These equation are 'sound', in the sense that they are respected by the string diagram interpretation of Subsection 7.1.2: $S_1 = S_2$ implies $[\![S_1]\!] = [\![S_2]\!]$. From now on we simplify the writing of string diagrams in the following way.

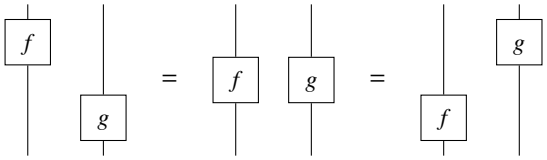
Convention 7.2.1. 1 We will write the arrows ↑ on the wires of string diagrams simply as | and assume that the flow is allways upwards in a string diagram, from bottom to top. Thus, even though we are not writing edges as arrows, we still consider string diagrams as directed graphs.

2 We drop the types of wires when they are clear from the context. Thus, wires in string diagrams are always typed, but we do not always write these types explicitly.

3 We become a bit sloppy about writing the interpretation function $[\![-]\!]$ for string diagrams explicitly. Sometimes we'll say "the interpretation of S" instead of just $[\![S]\!]$, but also we sometimes simply write a string diagram, where we mean its interpretation as a channel. This can be confusing at first, so we will make it explicit when we start blurring the distinction between syntax and semantics.

Shift equations

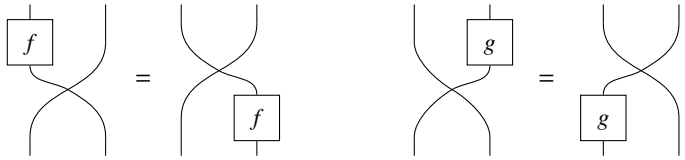
String diagrams allow quite a bit of topological freedom, for instance, in the sense that parallel boxes may be shifted up and down, as expressed by the following equations.



Semantically, this is justified by the equation:

$$([\![f]\!] \otimes id) \circ (id \otimes [\![g]\!]) = ([\![f]\!] \otimes [\![g]\!]) = (id \otimes [\![g]\!]) \circ ([\![f]\!] \otimes id).$$

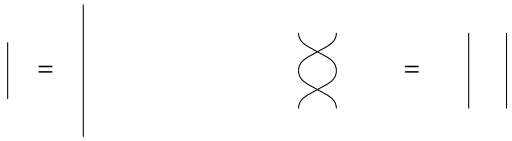
Similarly, boxes may be pulled trough swaps:



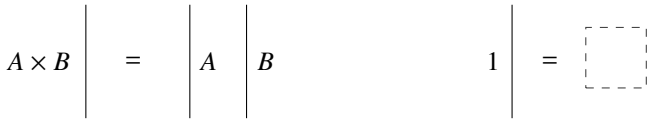
These four equations do not only hold for boxes, but also for $\bar{\top}$ and $\bar{\bot}$, since we consider them simply as boxes but with a special notation.

Equations for wires

Wires can be stretched arbitrarily in length, and successive swaps cancel each other out:

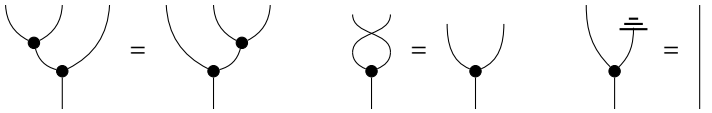


Wires of product type correspond to parallel wires, and wires of the empty product type 1 do nothing — as represented by the empty dashed box.



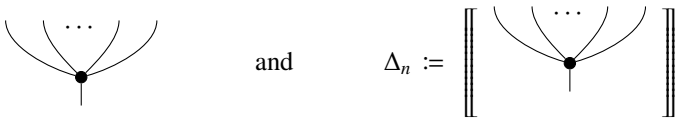
Equations for copying

The copy string diagram  also satisfies a number of equations, expressing associativity, commutativity and a co-unit property. Abstractly this makes a comonoid. This can be expressed graphically by:



(7.5)

Because copying is associative, it does not matter which order of copying we use in successive copying. Therefore we feel free to write such n-ary copying as coming from a single node $\bullet$, in:

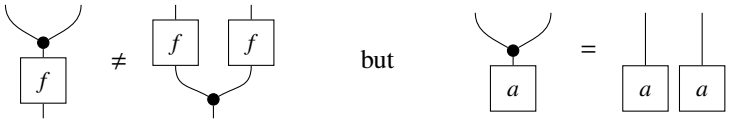


(7.6)

Copying of products $A \times B$ and of the empty product 1 comes with the following equations.



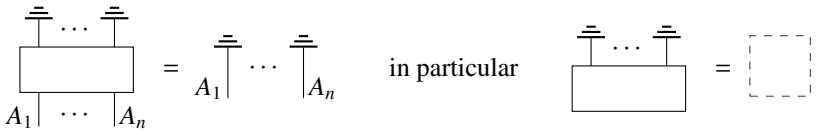
We should be careful that boxes can in general *not* be "pulled through" copiers, as expressed on the left below (see also Exercise 2.5.10).



An exception to this is when copying follows a point state $1|a\rangle$ for $a \in A$, as above, on the right. The above equation on the left does hold for deterministic channels f, see Exercise 2.5.12.

Equations for discarding and for uniform and point states

The most important discarding equation is the one immediately below, saying that discarding after a box/channel is the same as discarding of the inputs altogether. We refer to this as: boxes are *unital*[1].



Discarding only some (not all) of the outgoing wires amounts to marginalisation.

Discarding a uniform state also erases everything. This may be seen as a special case of the above equation on the right, but still we like to formulate it explicitly, below on the left. The other equations deal with trivial cases.



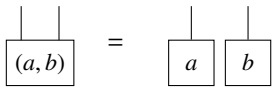
For product wires we have the following discard equations, see Exercise 7.1.6.



Similarly we have the following uniform state equations.

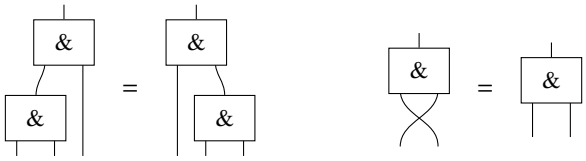


For point states we have for elements $a \in A$ and $b \in B$,



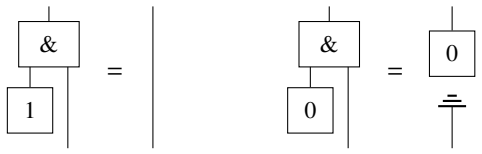
Equations for logical operations

As is well-known, conjunction & is associative and commutative:

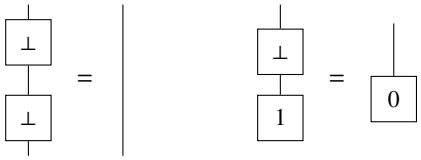


[1] In a quantum setting one uses 'causal' instead of 'unital', see *e.g.* [25].

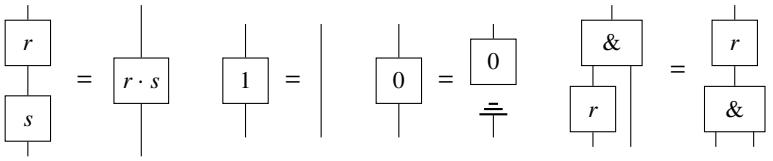
Moreover, conjunction has truth and falsity as unit and zero elements:



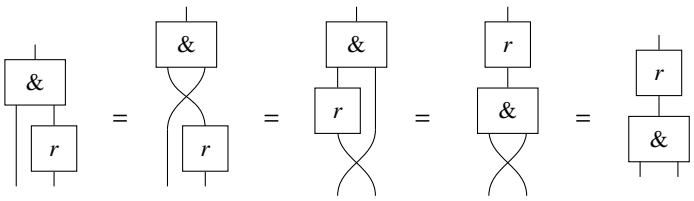
Orthosupplement $(-)^{\perp}$ is idempotent and turns 0 to 1 (and vice-versa):



Finally we look at the rules for scalar multiplication for predicates:



These equations be used for diagrammatic equational reasoning. For instance we can now derive a variation on the last equation, where scalar multiplication on the second input of conjunction can be done after the conjunction:

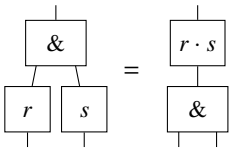


All of the above equations between string diagrams are sound, in the sense that they hold under all interpretations of string diagrams — as described at the end of the previous section. There are also completeness results in this area, but they require a deeper categorical analysis. For a systematic account of string diagrams we refer to the overview paper [149]. In this book we use string diagrams as convenient, intuitive notation with a precise semantics.

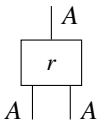
Exercises

7.2.1 Give a concrete description of the n-ary copy channel $\Delta_n \colon A \nrightarrow A^n$ in (7.6).

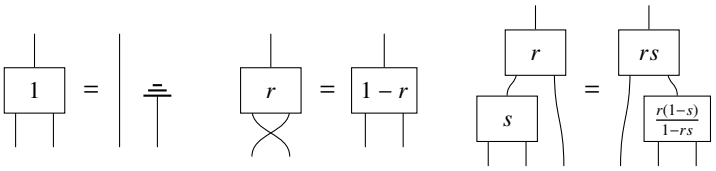
7.2.2 Prove by diagrammatic equational reasoning that:



7.2.3 Consider for $r \in [0,1]$ the convex combination channel $cc(r)\colon A \times A \nrightarrow A$ given by $cc(r)(a,a') = r|a\rangle + (1-r)|a'\rangle$. It can be used as interpretation of an additional convex combination box:

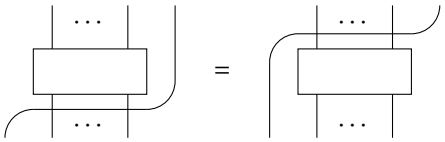


Prove that it satisfies the following 'barycentric' equations, due to [155].



For the last equation we must assume that $rs \neq 1$, *i.e.* not $r = s = 1$.

7.2.4 Prove that:

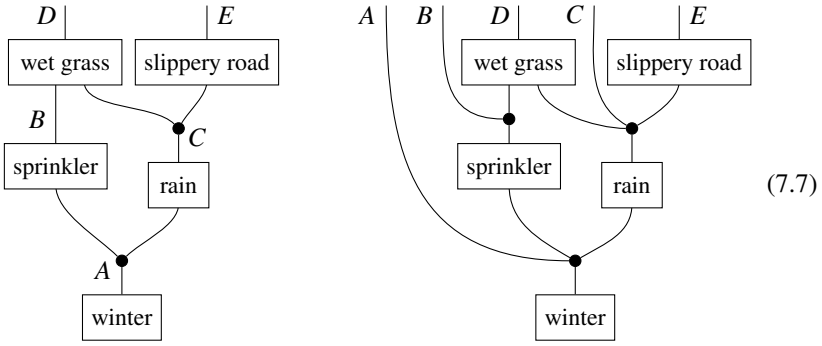


What does this amount to when there are no input (or no output) wires?

7.3 Accessibility and joint states

Intuitively, a string diagram is called *accessible* if all its internal connections are also accessible externally. This property will be defined formally below, but it is best illustrated via an example. Below on the left we see the wetness Bayesian network as string diagram, like in (7.3) — but with lines instead of arrows. On the right is an accessible version of this string diagram, with

(single) external connections to all its internal wires.



(7.7)

The non-accessible parts of a string diagram are often called *latent* or *unobserved*.

We are interested in such accessible versions because they lead in an easy way to joint states (and vice-versa, see below): the joint state for the wetness Bayesian network is the interpretation of the above accessible string diagram on the right, giving a distribution on the product set $A \times B \times D \times C \times E$. Such joint distributions are conceptually important, but they are often impractical to work with because their size quickly grows out of hand. For instance, the joint distribution ω for the above wetness network is (with zero-probability terms omitted, as usual):

$$\begin{aligned}
&\tfrac{399}{6250}|a,b,d,c,e\rangle + \tfrac{171}{6250}|a,b,d,c,e^\perp\rangle + \tfrac{27}{1250}|a,b,d,c^\perp,e^\perp\rangle \\
&+ \tfrac{21}{6250}|a,b,d^\perp,c,e\rangle + \tfrac{9}{6250}|a,b,d^\perp,c,e^\perp\rangle + \tfrac{3}{1250}|a,b,d^\perp,c^\perp,e^\perp\rangle \\
&+ \tfrac{672}{3125}|a,b^\perp,d,c,e\rangle + \tfrac{288}{3125}|a,b^\perp,d,c,e^\perp\rangle + \tfrac{168}{3125}|a,b^\perp,d^\perp,c,e\rangle \\
&+ \tfrac{72}{3125}|a,b^\perp,d^\perp,c,e^\perp\rangle + \tfrac{12}{125}|a,b^\perp,d^\perp,c^\perp,e^\perp\rangle + \tfrac{399}{20000}|a^\perp,b,d,c,e\rangle \\
&+ \tfrac{171}{20000}|a^\perp,b,d,c,e^\perp\rangle + \tfrac{243}{1000}|a^\perp,b,d,c^\perp,e^\perp\rangle + \tfrac{21}{20000}|a^\perp,b,d^\perp,c,e\rangle \\
&+ \tfrac{9}{20000}|a^\perp,b,d^\perp,c,e^\perp\rangle + \tfrac{27}{1000}|a^\perp,b,d^\perp,c^\perp,e^\perp\rangle + \tfrac{7}{1250}|a^\perp,b^\perp,d,c,e\rangle \\
&+ \tfrac{3}{1250}|a^\perp,b^\perp,d,c,e^\perp\rangle + \tfrac{7}{5000}|a^\perp,b^\perp,d^\perp,c,e\rangle + \tfrac{3}{5000}|a^\perp,b^\perp,d^\perp,c,e^\perp\rangle \\
&+ \tfrac{9}{100}|a^\perp,b^\perp,d^\perp,c^\perp,e^\perp\rangle
\end{aligned}$$

This distribution is obtained as outcome of the interpreted accessible graph in (7.7):

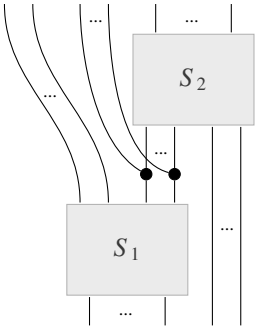
$$\omega = (id \otimes id \otimes wg \otimes id \otimes sr) \circ (id \otimes \Delta_2 \otimes \Delta_3) \circ (id \otimes sp \otimes sr) \circ \Delta_3 \circ wi.$$

Recall that Δ_n is written for the n-ary copy channel, see (7.6).

Later on in this section we use our new insights into accessible string diagrams to re-examine the relation between crossover inference on joint states

and channel-based inference, see Corollary 6.3.13. But we start with a more precise description.

Definition 7.3.1. 1 An *accessible string diagram* is defined inductively via the string diagram definitions steps (1) – (3) from Section 7.1, with sequential composition diagram (7.2) in step (4) extended with copiers on the internal wires:



(7.8)

In this way each 'internal' wire between string diagrams S_1 and S_2 is 'externally' accessible.

2 Each string diagram S can be made accessible by replacing each occurrence (7.2) of step (4) in its construction with the above modified step (7.8). We shall write $\overline{S}$ for a choice of accessible version of S.

3 A *joint state* or *joint distribution* associated with a Bayesian network $G \in SD(\Sigma)$ is the interpretation $[\![\overline{G}]\!]$ of an accessible version $\overline{G} \in SD(\Sigma)$ of the string diagram G. It re-uses G's interpretation of its boxes.

There are several subtleties related to the last two points.

Remark 7.3.2. 1 The above definition of accessible string diagram allows that there are multiple external connections to an internal wire, for instance if one of the internal wires between S_1 and S_2 in (7.8) is made accessible inside S_2, possibly because it is one of the outgoing wires of S_2. Such multiple accessibility is unproblematic, but in most cases we prefer 'minimal' accessibility via single wires, as in (7.7) on the right. Anyway, we can always purge unnecessary outgoing wires via discard $\top$.

2 The order of the outgoing wires in an accessible graph involves some level of arbitrairiness. For instance, in the accessible wetness graph in (7.7) one could also imagine putting C to the left of D at the top of the diagram. This would involve some additional crossing of wires. The result would be essentially the same. The resulting joint state, obtained by interpretation,

would then be in $\mathcal{D}(A \times B \times C \times D \times E)$. These differences are immaterial, but do require some care if one performs marginalisation.

Thus, strictly speaking, the accessible version $\overline{S}$ of a string diagram S is determined only up to permutation of outgoing wires. In concrete cases we try to write accessible string diagrams in such a way that the number of crossings of wires is minimal.

With these points in mind we allow ourselves to speak about *the* joint state associated with a Bayesian network, in which duplicate outgoing wires are deleted. This joint state is determined up-to permutation of outgoing wires.

In the crossover inference corollary 6.3.13 we have seen how component-wise updates on a joint state correspond to channel-based inference. This theorem involved a binary state and only one channel. Now that we have a better handle on joint states via string diagrams, we can extend this correspondence to n-ary joint states and multiple channels. This is a theme that will be pursued further in this chapter. At this stage we only illustrate how this works for the wetness Bayesian network.

Example 7.3.3. In Exercise 6.5.1 we have formulated two inference questions for the wetness Bayesian network, namely:

1 What is the updated sprinkler distribution, given evidence of a slippery road? Using channel-based inference it is computed as:

$$sp \gg\!= \left(wi\big|_{ra \prec\!\ll (sr \prec\!\ll \mathbf{1}_e)}\right) = \tfrac{63}{260}|b\rangle + \tfrac{197}{260}|b^\perp\rangle.$$

2 What is the updated wet grass distribution, given evidence of a slippery road? It's:

$$wg \gg\!= \Big(((sp \otimes ra) \gg\!= (\Delta \gg\!= wi))\big|_{\mathbf{1}\otimes(sr \prec\!\ll \mathbf{1}_e)}\Big) = \tfrac{4349}{5200}|d\rangle + \tfrac{851}{5200}|d^\perp\rangle.$$

Via (multiple applications) of crossover inference, see Corollary 6.3.13, these same updated probabilities can be obtained via the joint state $\omega \in \mathcal{D}(A \times B \times D \times C \times E)$ of the wetness Bayesian network, as mentioned at the beginning of this section.

Concretely, this works as follows. In the above first case we have (point) evidence $\mathbf{1}_e \colon E \to [0,1]$ on the set $E = \{e, e^\perp\}$. It is weakened (extended) to evidence $\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_e$ on the product $A\times B\times D\times C\times E$. Then it can be used to update the joint state ω. The required outcome then appears by marginalising on the sprinkler set B in second position, as in:

$$\begin{aligned}\omega\big|_{\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_e}[0,1,0,0,0] &= \tfrac{63}{260}|b\rangle + \tfrac{197}{260}|b^\perp\rangle \\ &\approx 0.2423|b\rangle + 0.7577|b^\perp\rangle.\end{aligned}$$

For the second inference question we use the same evidence, but we now marginalise on the wet grass set D, in third position:

$$\begin{aligned}\omega\big|_{\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_e}[0,0,1,0,0] &= \tfrac{4349}{5200}|d\rangle + \tfrac{851}{5200}|d^\perp\rangle \\ &\approx 0.8363|d\rangle + 0.1637|d^\perp\rangle.\end{aligned}$$

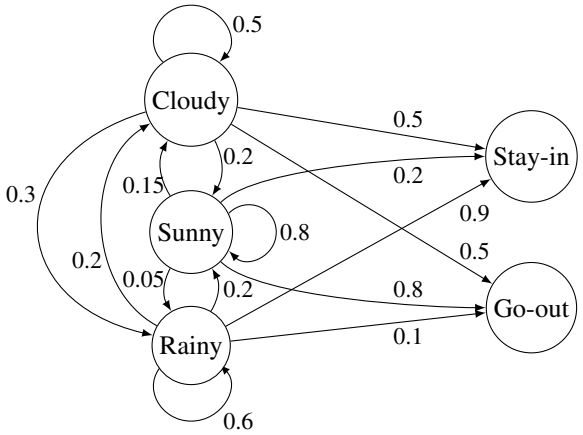
We see that crossover inference is easier to express than channel-based inference, since we do not have to form the more complicated expressions with $\gg=$ and $=\ll$ in the above two points. Instead, we just have to update and marginalise at the right positions. Thus, it is easier from a mathematical perspective. But from a computational perspective crossover inference is more complicated, since these joint states grow exponentially in size (in the number of nodes in the graph). This topic will be continued in Section 7.10.

Exercises

7.3.1 Draw an accessible verion of the Asia string diagram in (7.3). Write also the corresponding composition of channels that produces the associated joint state.

7.4 Hidden Markov models

Let's start with a graphical description of an example of what is called a hidden Markov model:



(7.9)

This model has three 'hidden' elements, namely Cloudy, Sunny, and Rainy, representing the weather condition on a particular day. There are 'temporal' transitions with associated probabilities between these elements, as indicated

by the labeled arrows. For instance, if it is cloudy today, then there is a 50% chance that it will be cloudy again tomorrow. There are also two 'visible' elements on the right: Stay-in, and Stay-out, describing two possible actions of a person, depending on the weather condition. There are transitions with probabilities from the hidden to the visible elements. The idea is that with every time step a transition is made between hidden elements, resulting in a visible outcome. Such steps may be repeated for a finite number of times — or even forever. The interaction between what is hidden and what can be observed is a key element of hidden Markov models. For instance, one may ask: given a certain initial state, how likely is it to see a consecutive sequence of the four visible elements: Stay-in, Stay-in, Go-out, Stay-in?

Hidden Markov models are simple statistical models that have many applications in temporal pattern recognition, in speech, handwriting or gestures, but also in robotics and in biological sequences. This section will briefly look into hidden Markov models, using the notation and terminology of channels and string diagrams. Indeed, a hidden Markov model can be defined easily in terms of channels and forward and backward transformation of states and observables. In addition, conditioning of states by observables can be used to formulate and answer elementary questions about hidden Markov models. Learning for Markov models will be described separately in Sections **??** and **??**. Markov models are examples of probabilistic automata. Such automata will be studied separately, in Chapter **??**.

Definition 7.4.1. A *Markov model* (or a *Markov chain*) is given by a set X of 'internal positions', typically finite, with a 'transition' channel $t\colon X \nrightarrow X$ and an initial state/distribution $\sigma \in \mathcal{D}(X)$.

A *hidden Markov model*, often abbreviated as HMM, is a Markov model, as just described, with an additional 'emission' channel $e\colon X \nrightarrow Y$, where Y is a set of 'outputs'.

In the above illustration (7.9) we have as sets of positions and outputs:

$$X = \{\text{Cloudy}, \text{Sunny}, \text{Rainy}\} \qquad Y = \{\text{Stay-in}, \text{Go-out}\},$$

with transition channel $t\colon X \nrightarrow X$,

$$\begin{aligned} t(\text{Cloudy}) &= 0.5\big|\text{Cloudy}\big\rangle + 0.2\big|\text{Sunny}\big\rangle + 0.3\big|\text{Rainy}\big\rangle \\ t(\text{Sunny}) &= 0.15\big|\text{Cloudy}\big\rangle + 0.8\big|\text{Sunny}\big\rangle + 0.05\big|\text{Rainy}\big\rangle \\ t(\text{Rainy}) &= 0.2\big|\text{Cloudy}\big\rangle + 0.2\big|\text{Sunny}\big\rangle + 0.6\big|\text{Rainy}\big\rangle, \end{aligned}$$

and emission channel $e\colon X \nrightarrow Y$,

$$\begin{aligned} e(\text{Cloudy}) &= 0.5\big|\text{Stay-in}\rangle + 0.5\big|\text{Go-out}\rangle \\ e(\text{Sunny}) &= 0.2\big|\text{Stay-in}\rangle + 0.8\big|\text{Go-out}\rangle \\ e(\text{Rainy}) &= 0.9\big|\text{Stay-in}\rangle + 0.1\big|\text{Go-out}\rangle. \end{aligned}$$

An initial state is missing in the picture (7.9).

In the literature on Markov models, the elements of the set X are often called states. This clashes with the terminology in this book, since we use 'state' as synonym for 'distribution'. So, here we call $\sigma \in \mathcal{D}(X)$ an (initial) state, and we call elements of X (internal) positions. At the same time we may call X the sample space. The transition channel $t\colon X \nrightarrow X$ is an endo-channel on X, that is a channel from X to itself. As a function, it is of the form $t\colon X \to \mathcal{D}(X)$; it is an instance of a *coalgebra*, that is, a map of the form $A \to F(A)$ for a functor F, see Section **??** for more information.

In a Markov chain/model one can iteratitively compute successor states. For an initial state σ and transition channel t one can form successor states via state transformation:

$$\begin{array}{l} \sigma \\ t \gg\!= \sigma \\ t \gg\!= (t \gg\!= \sigma) \;=\; (t \mathbin{\odot} t) \gg\!= \sigma \qquad \text{where} \qquad t^n = \begin{cases} \mathit{unit} & \text{if } n = 0 \\ t \mathbin{\odot} t^{n-1} & \text{if } n > 0. \end{cases} \\ \quad \vdots \\ t^n \gg\!= \sigma \end{array}$$

In these transitions the state at stage $n+1$ only depends on the state at stage n: in order to predict a future step, all we need is the immediate predecessor state. This makes HMMs relatively easy dynamical models. Multi-stage dependencies can be handled as well, by enlarging the sample space, see Exercise 7.4.6 below.

One interesting problem in the area of Markov chains is to find a 'stationary' state σ_∞ with $t \gg\!= \sigma_\infty = \sigma_\infty$, see Exercise 7.4.2 for an illustration, and also Exercise 2.5.16 for a sufficient condition.

Here we are more interested in hidden Markov models $1 \overset{\sigma}{\nrightarrow} X \overset{t}{\nrightarrow} X \overset{e}{\nrightarrow} Y$. The elements of the set Y are observable — and hence sometimes called *signals* — whereas the elements of X are hidden. Thus, many questions related to hidden Markov models concentrate on what one can learn about X via Y, in a finite number of steps. Hidden Markov models are examples of models with *latent* variables.

We briefly discuss some basic issues related to HMMs in separate subsec-

tions. A recurring theme is the relationship 'joint' and 'sequential' formulations.

7.4.1 Validity in hidden Markov models

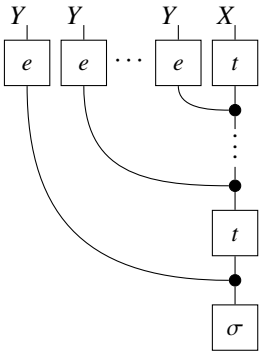
The first question that we like to address is: given a sequence of observables, what is their probability (validity) in a HMM? Standardly in the literature, one only looks at the probability of a sequence of point observations (elements), but here we use a more general approach. After all, one may not be certain about observing a specific point at a particular state, or some point observations may be missing; in the latter case one may wish to replace them by a constant (uniform) observation.

We proceed by defining validity of the sequence of observables in a joint state first; subsequently we look at (standard) algorithms for computing these validities efficiently. We thus start by defining a relevant joint state.

We fix a HMM $1 \overset{\sigma}{\oto} X \overset{t}{\oto} X \overset{e}{\oto} Y$. For each $n \in \mathbb{N}$ a channel $\langle e, t\rangle_n \colon X \oto Y^n \times X$ is defined in the following manner:

$$\begin{aligned} \langle e, t\rangle_0 &:= \Big(X \overset{id}{\longrightarrow\!\!\circ} X \cong 1 \times X \cong Y^0 \times X\Big) \\ \langle e, t\rangle_{n+1} &:= \Big(X \overset{\langle e,t\rangle_n}{\longrightarrow\!\!\circ} Y^n \times X \overset{id^n \otimes \langle e,t\rangle}{\longrightarrow\!\!\circ} Y^n \times (Y \times X) \cong Y^{n+1} \times X\Big). \end{aligned} \tag{7.10}$$

We recall that the tuple $\langle e, t\rangle$ of channels is $(e \otimes t) \circ \Delta$, see Definition 2.5.6 (1). With these tuples we can form a joint state $\langle e, t\rangle_n \gg\!= \sigma \in \mathcal{D}(Y^n \times X)$. As an (interpreted) string diagram it looks as follows.



(7.11)

We consider the combined likelihood of a sequence of observables on the set Y in a hidden Markov model. In the literature these observables are typically point predicates $\mathbf{1}_y \colon Y \to [0, 1]$, for $y \in Y$, but, as mentioned, here we allow more general observables $Y \to \mathbb{R}$.

Definition 7.4.2. Let $\mathcal{H} = (1 \overset{\sigma}{\dashrightarrow} X \overset{t}{\dashrightarrow} X \overset{e}{\dashrightarrow} Y)$ be a hidden Markov model and let $\vec{p} = p_1, \ldots, p_n$ be a list of observables on Y. The validity $\mathcal{H} \models \vec{p}$ of this sequence $\vec{p}$ in the model $\mathcal{H}$ is defined via the tuples (7.10) as:

$$\begin{aligned} \mathcal{H} \models \vec{p} &:= (\langle e, t\rangle_n \gg\!= \sigma)[1, \ldots, 1, 0] \models p_1 \otimes \cdots \otimes p_n \\ &\overset{(4.8)}{=} \langle e, t\rangle_n \gg\!= \sigma \models p_1 \otimes \cdots \otimes p_n \otimes \mathbf{1} \\ &\overset{(4.9)}{=} \sigma \models \langle e, t\rangle_n =\!\ll (p_1 \otimes \cdots \otimes p_n \otimes \mathbf{1}). \end{aligned} \tag{7.12}$$

The marginalisation mask $[1, \ldots, 1, 0]$ contains n times the number 1. It ensures that the X outcome in (7.11) is discarded.

We describe an alternative way to formulate this validity without using the (big) joint state on $Y^n \times X$. It forms the essence of the classical 'forward' and 'backward' algorithms for validity in HMMs, see *e.g.* [14, 144] or [96, App. A]. An alternative algorithm is described in Exercise 7.4.4.

Proposition 7.4.3. *The HMM-validity* (7.12) *can be computed as:*

$$\begin{aligned} \mathcal{H} \models \vec{p} \;=\; \sigma \models\; & (e =\!\ll p_1) \;\&\; \\ & t =\!\ll (\, (e =\!\ll p_2) \;\&\; \\ & \qquad t =\!\ll (\, (e =\!\ll p_3) \;\&\; \cdots \\ & \qquad\qquad t =\!\ll (e =\!\ll p_n) \cdots \,)). \end{aligned} \tag{7.13}$$

This validity can be calculated recursively in forward manner as:

$$\sigma \models \alpha(\vec{p}) \qquad \textit{where} \qquad \begin{cases} \alpha([q]) = e =\!\ll q \\ \alpha([q] +\!\!+ \vec{q}) = (e =\!\ll q) \;\&\; (t =\!\ll \alpha(\vec{q})). \end{cases}$$

Alternatively, this validity can be calculated recursively in backward manner as:

$$\sigma \models \beta(\vec{p}, \mathbf{1}) \quad \textit{where} \quad \begin{cases} \beta([q]) = q \\ \beta(\vec{q} +\!\!+ [q_n, q_{n+1}]) = \beta(\vec{q} +\!\!+ [(e =\!\ll q_n) \;\&\; (t =\!\ll q_{n+1})]). \end{cases}$$

Proof. We first prove, by induction on $n \geq 1$ that for observables p_i on Y and q on X one has:

$$\begin{aligned} &\langle e, t\rangle_n =\!\ll (p_1 \otimes \cdots \otimes p_n \otimes q) \\ &= (e =\!\ll p_1) \;\&\; t =\!\ll ((e =\!\ll p_2) \;\&\; t =\!\ll (\cdots t =\!\ll ((e =\!\ll p_n) \;\&\; t =\!\ll q) \cdots)) \end{aligned} \tag{$*$}$$

The base case $n = 1$ is easy:

$$\begin{aligned} \langle e, t\rangle_1 =\!\ll (p_1 \otimes q) &= \Delta =\!\ll ((e \otimes t) =\!\ll (p_1 \otimes q)) \\ &= \Delta =\!\ll ((e =\!\ll p_1) \otimes (t =\!\ll q)) \qquad \text{by Exercise 4.3.8} \\ &= (e =\!\ll p_1) \;\&\; (t =\!\ll q) \qquad \text{by Exercise 4.3.7.} \end{aligned}$$

For the induction step we reason as follows.

$$\begin{aligned}
&\langle e,t\rangle_{n+1} \mathrel{=\!\!\ll} (p_1 \otimes \cdots \otimes p_n \otimes p_{n+1} \otimes q) \\
&\quad = \langle e,t\rangle_n \mathrel{=\!\!\ll} ((id^n \otimes \langle t,e\rangle) \mathrel{=\!\!\ll} (p_1 \otimes \cdots \otimes p_n \otimes p_{n+1} \otimes q)) \\
&\quad = \langle e,t\rangle_n \mathrel{=\!\!\ll} (p_1 \otimes \cdots \otimes p_n \otimes (\langle e,t\rangle \mathrel{=\!\!\ll} (p_{n+1} \otimes q)) \\
&\quad = \langle e,t\rangle_n \mathrel{=\!\!\ll} (p_1 \otimes \cdots \otimes p_n \otimes ((e \mathrel{=\!\!\ll} p_{n+1}) \mathbin{\&} (t \mathrel{=\!\!\ll} q))) \qquad \text{as just shown} \\
&\overset{\text{(IH)}}{=} (e \mathrel{=\!\!\ll} p_1) \mathbin{\&} t \mathrel{=\!\!\ll} ((e \mathrel{=\!\!\ll} p_2) \mathbin{\&} t \mathrel{=\!\!\ll} (\cdots \\
&\qquad t \mathrel{=\!\!\ll} ((e \mathrel{=\!\!\ll} p_n) \mathbin{\&} t \mathrel{=\!\!\ll} ((e \mathrel{=\!\!\ll} p_{n+1}) \mathbin{\&} (t \mathrel{=\!\!\ll} q))) \cdots)).
\end{aligned}$$

We can now prove Equation (7.13):

$$\begin{aligned}
&\mathcal{H} \models \vec{p} \\
&\overset{(7.12)}{=} \sigma \models \langle e,t\rangle_n \mathrel{=\!\!\ll} (p_1 \otimes \cdots \otimes p_n \otimes \mathbf{1}) \\
&\overset{(*)}{=} (e \mathrel{=\!\!\ll} p_1) \mathbin{\&} t \mathrel{=\!\!\ll} ((e \mathrel{=\!\!\ll} p_2) \mathbin{\&} t \mathrel{=\!\!\ll} (\cdots t \mathrel{=\!\!\ll} ((e \mathrel{=\!\!\ll} p_n) \mathbin{\&} t \mathrel{=\!\!\ll} \mathbf{1}) \cdots)) \\
&= (e \mathrel{=\!\!\ll} p_1) \mathbin{\&} t \mathrel{=\!\!\ll} ((e \mathrel{=\!\!\ll} p_2) \mathbin{\&} t \mathrel{=\!\!\ll} (\cdots t \mathrel{=\!\!\ll} (e \mathrel{=\!\!\ll} p_n) \cdots)). \qquad \text{❑}
\end{aligned}$$

7.4.2 Filtering

Given a sequence $\vec{p}$ of factors, one can compute their validity $\mathcal{H} \models \vec{p}$ in a HMM $\mathcal{H}$, as described above. But we can also use these factors to 'guide' the evolution of the HMM. At each state i the factor p_i is used to update the current state, via backward inference. The new state is then moved forward via the transition function. This process is called *filtering*, after the Kalman filter from the 1960s that is used for instance in trajectory optimisation in navigation and in rocket control (*e.g.* for the Apollo program). The system can evolve autonomously via its transition function, but observations at regular intervals can update (correct) the current state.

Definition 7.4.4. Let $\mathcal{H} = (1 \overset{\sigma}{\multimap} X \overset{t}{\multimap} X \overset{e}{\multimap} Y)$ be a hidden Markov model and let $\vec{p} = p_1, \ldots, p_n$ be a list of factors on Y. It gives rise to the *filtered* sequence of states $\sigma_1, \sigma_1, \ldots, \sigma_{n+1} \in \mathcal{D}(X)$ following the observe-update-proceed principle:

$$\sigma_1 \coloneqq \sigma \qquad \text{and} \qquad \sigma_{i+1} \coloneqq t \gg\!= (\sigma_i\big|_{e \mathrel{=\!\!\ll} p_i}).$$

In the terminology of Definition 6.3.1, the definition of the state σ_{i+1} involves both forward and backward inference. Below we show that the final state σ_{n+1} in the filtered sequence can also be obtained via crossover inference on a joint state, obtained via the tuple channels (7.10). This fact gives a theoretical justification, but is of little practical relevance — since joint states quickly become too big to handle.

Proposition 7.4.5. *In the context of Definition 7.4.4,*

$$\sigma_{n+1} \;=\; (\langle e,t\rangle_n \gg\!= \sigma)\Big|_{p_1\otimes\cdots\otimes p_n\otimes \mathbf{1}}[0,\ldots,0,1]$$

The marginalisation mask $[0,\ldots,0,1]$ *has n zero's.*

Proof. By induction on $n \geq 1$. The base case with $\langle e,t\rangle_1 = \langle e,t\rangle$ is handled as follows.

$$\begin{aligned}
&(\langle e,t\rangle \gg\!= \sigma)\Big|_{p_1\otimes\mathbf{1}}[0,1] \\
&= \pi_2 \gg\!= (\langle e|_{p_1}, t\rangle \gg\!= (\sigma|_{\langle e,t\rangle \prec\!\ll (p_1\otimes\mathbf{1})})) \quad \text{by Corollary 6.3.12 (2)} \\
&= t \gg\!= (\sigma|_{e \prec\!\ll p_1}) \\
&= \sigma_2.
\end{aligned}$$

The induction step requires a bit more work:

$$\begin{aligned}
&(\langle e,t\rangle_{n+1} \gg\!= \sigma)\Big|_{p_1\otimes\cdots\otimes p_n\otimes p_{n+1}\otimes\mathbf{1}}[0,\ldots,0,0,1] \\
&= \pi_{n+2} \gg\!= \Big((id^n \otimes \langle e,t\rangle) \gg\!= (\langle e,t\rangle_n \gg\!= \sigma))\Big|_{p_1\otimes\cdots\otimes p_n\otimes p_{n+1}\otimes\mathbf{1}}\Big) \\
&\overset{(6.2)}{=} \pi_{n+2} \gg\!= \Big((id^n \otimes \langle e,t\rangle)\Big|_{p_1\otimes\cdots\otimes p_n\otimes p_{n+1}\otimes\mathbf{1}} \\
&\qquad\qquad \gg\!= (\langle e,t\rangle_n \gg\!= \sigma)\Big|_{(id^n\otimes\langle e,t\rangle) \prec\!\ll (p_1\otimes\cdots\otimes p_n\otimes p_{n+1}\otimes\mathbf{1})}\Big) \\
&= (\pi_{n+2} \gg\!= (id^n \otimes \langle e|_{p_{n+1},t}\rangle)) \gg\!= \Big((\langle e,t\rangle_n \gg\!= \sigma)\Big|_{p_1\otimes\cdots\otimes p_n\otimes((e\prec\!\ll p_{n+1})\otimes(t\prec\!\ll\mathbf{1}))}\Big) \\
&= (t \circ \pi_{n+1}) \gg\!= \Big((\langle e,t\rangle_n \gg\!= \sigma)\Big|_{p_1\otimes\cdots\otimes p_n\otimes(e\prec\!\ll p_{n+1})}\Big) \\
&= t \gg\!= \Big(((\langle e,t\rangle_n \gg\!= \sigma)\Big|_{p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}}\Big|_{\mathbf{1}\otimes\cdots\otimes\mathbf{1}\otimes(e\prec\!\ll p_{n+1})})[0,\ldots,0,1]\Big) \\
&\qquad \text{by Lemma 6.1.6 (3) and Lemma 4.2.9 (1)} \\
&= t \gg\!= \Big(((\langle e,t\rangle_n \gg\!= \sigma)\Big|_{p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}})[0,\ldots,0,1]\Big|_{e\prec\!\ll p_{n+1}}\Big) \\
&\qquad \text{by Lemma 6.1.6 (6)} \\
&\overset{(IH)}{=} t \gg\!= (\sigma_{n+1}\Big|_{e\prec\!\ll p_{n+1}}) \\
&= \sigma_{n+2}.
\end{aligned}$$

❑

The next consequence of the previous proposition may be understood as Bayes' rule for HMMs — or more accurately, the product rule for HMMs, see Proposition 6.1.3.

Corollary 7.4.6. *Still in the context of Definition 7.4.4, let q be a predicate on X. Its validity in the final state in the sequence* $\sigma_1,\ldots,\sigma_{n+1}$*, filtered by factors* $p_1,\ldots,p_n$*, is given by:*

$$\sigma_{n+1} \models q \;=\; \frac{\langle e,t\rangle_n \gg\!= \sigma \models p_1 \otimes\cdots\otimes p_n \otimes q}{\mathcal{H} \models \vec{p}}.$$

Proof. This follows from Bayes' rule, in Proposition 6.1.3 (1):

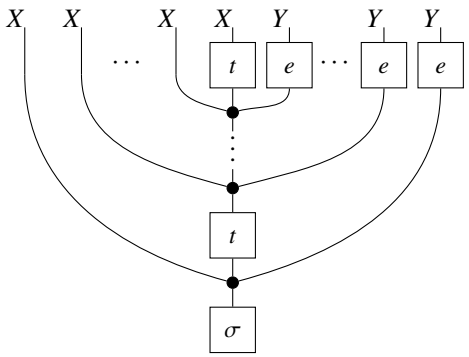
$$\begin{aligned}
\sigma_{n+1} \models q &= (\langle e, t\rangle_n \gg\!= \sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes \mathbf{1}}[0,\ldots,0,1] \models q \\
&= (\langle e, t\rangle_n \gg\!= \sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes \mathbf{1}} \models \mathbf{1}\otimes\cdots\otimes\mathbf{1}\otimes q \qquad \text{by (4.8)} \\
&= \frac{(\langle e, t\rangle_n \gg\!= \sigma) \models (p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}) \mathbin{\&} (\mathbf{1}\otimes\cdots\otimes\mathbf{1}\otimes q)}{(\langle e, t\rangle_n \gg\!= \sigma) \models p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}} \\
&= \frac{(\langle e, t\rangle_n \gg\!= \sigma) \models p_1\otimes\cdots\otimes p_n\otimes q}{\mathcal{H} \models \vec{p}}.
\end{aligned}$$

The last equation uses Lemma 4.2.9 (1) and Definition 7.4.2. ❑

7.4.3 Finding the most likely sequence of hidden elements

We briefly consider one more question for HMMs: given a sequence of observations, what is the most likely path of internal positions that produces these observations? This is also known as the decoding problem. It explicitly asks for the most likely *path*, and not for the most likely individual position at each stage. Thus it involves taking the argmax of a joint state. We concentrate on giving a definition of the solution, and only sketch how to obtain it efficiently.

We start by constructing the state below, as intepreted string diagram, for a given HMM $1 \overset{\sigma}{\dashrightarrow} X \overset{t}{\dashrightarrow} X \overset{e}{\dashrightarrow} Y$.



(7.14)

We proceed, much like in the beginning of this section, this time not using tuples but triples. We define the above state as element of the subset:

$$\langle id, t, e\rangle_n \gg\!= \sigma \qquad \text{for} \qquad X \xrightarrow{\langle id,t,e\rangle_n} X^n \times X \times Y^n$$

These maps $\langle id, t, e\rangle_n$ are obtained by induction on n:

$$\begin{aligned}\langle id, t, e\rangle_0 &:= \left(X \xrightarrow{id} X \cong 1 \times X \times 1 \cong X^0 \times X \times Y^0\right)\\ \langle id, t, e\rangle_{n+1} &:= \left(X \xrightarrow{\langle id,t,e\rangle_n} X^n \times X \times Y^n \xrightarrow{id^n \otimes \langle id,t,e\rangle \otimes id^n} X^n \times (X \times X \times Y) \times Y^n \right.\\ &\qquad\qquad \left. \cong X^{n+1} \times X \times Y^{n+1}\right)\end{aligned}$$

Definition 7.4.7. Let $1 \overset{\sigma}{\to} X \overset{t}{\to} X \overset{e}{\to} Y$ be a HMM and let $\vec{p} = p_1, \ldots, p_n$ be a sequence of factors on Y. Given these factors as successive observations, the most likely path of elements of the sample space X, as an n-tuple in X^n, is obtained as:

$$\operatorname{argmax}\left(\left(\langle id, t, e\rangle_n \gg\!= \sigma\right)\Big|_{\mathbf{1}\otimes\cdots\otimes\mathbf{1}\otimes p_n\otimes\cdots\otimes p_1}[1, \ldots, 1, 0, 0, \ldots, 0]\right). \tag{7.15}$$

This expression involves updating the joint state (7.14) with the factors p_i, in reversed order, and then taking its marginal so that the state with the first n outcomes in X remains. The argmax of the latter state gives the sequence in X^n that we are after.

The expression (7.15) is important conceptually, but not computationally. It is inefficient to compute, first of all because it involves a joint state that grows exponentionally in n, and secondly because the conditioning involves normalisation that is irrelevant when we take the argmax.

We given an impression of what (7.15) amounts to for $n = 3$. We first focus on the expression within $\operatorname{argmax}(-)$. The marginalisation and conditioning amount produce an state on $X \times X \times X$ of the form:

$$\begin{aligned}&\sum_{x,y_1,y_2,y_3} \frac{(\langle id, t, e\rangle_3 \gg\!= \sigma)(x_1, x_2, x_3, x, y_3, y_2, y_1) \cdot p_3(y_3) \cdot p_2(y_2) \cdot p_1(y_1)}{\langle id, t, e\rangle_3 \gg\!= \sigma \models \mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1} \otimes p_3 \otimes p_2 \otimes p_1}\\ &\sim \sum_{x,y_1,y_2,y_3} \sigma(x_1) \cdot t(x_1)(x_2) \cdot t(x_2)(x_3) \cdot t(x_3)(x)\\ &\qquad \cdot e(x_3)(y_3) \cdot e(x_2)(y_2) \cdot e(x_1)(y_1) \cdot p_3(y_3) \cdot p_2(y_2) \cdot p_1(y_1)\\ &= \sigma(x_1) \cdot \Big(\textstyle\sum_{y_1} e(x_1)(y_1) \cdot p_1(y_1)\Big) \cdot t(x_1)(x_2) \cdot \Big(\textstyle\sum_{y_2} e(x_2)(y_2) \cdot p_2(y_2)\Big)\\ &\qquad \cdot t(x_2)(x_3) \cdot \Big(\textstyle\sum_{y_3} e(x_3)(y_3) \cdot p_3(y_3)\Big) \cdot \Big(\textstyle\sum_x t(x_3)(x)\Big)\\ &= \sigma(x_1) \cdot (e \ll\!= p_1)(x_1) \cdot t(x_1)(x_2) \cdot (e \ll\!= p_2)(x_2) \cdot t(x_2)(x_3) \cdot (e \ll\!= p_3)(x_3).\end{aligned}$$

This approach is known as the Viterbi algorithm, see *e.g.* [14, 144, 96].

Exercises

7.4.1 Consider the HMM example 7.9 with initial state $\sigma = 1|\text{Cloudy}\rangle$.

1 Compute successive states $t^n \gg\!= \sigma$ for $n = 0, 1, 2, 3$.

2 Compute successive observations $e \gg\!= (t^n \gg\!= \sigma)$ for $n = 0, 1, 2, 3$.

3 Check that the validity of the sequence of (point-predicate) observations Go-out, Stay-in, Stay-in is 0.1674.

4 Show that filtering, as in Definition 7.4.4, with these same three (point) observations yields as final outcome:

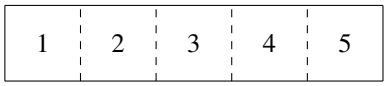
$$\begin{aligned}&\tfrac{1867}{6696}|\text{Cloudy}\rangle + \tfrac{347}{1395}|\text{Sunny}\rangle + \tfrac{15817}{33480}|\text{Rainy}\rangle \\ &\approx 0.2788|\text{Cloudy}\rangle + 0.2487|\text{Sunny}\rangle + 0.4724|\text{Rainy}\rangle.\end{aligned}$$

7.4.2 Consider the transition channel t associated with the HMM example 7.9. Check that in order to find a stationary state $\sigma_\infty = x|\text{Cloudy}\rangle + y|\text{Sunny}\rangle + z|\text{Rainy}\rangle$ one has to solve the equations:

$$\begin{aligned}x &= 0.5x + 0.15y + 0.2z \\ y &= 0.2x + 0.8y + 0.2z \\ z &= 0.3x + 0.05y + 0.6z\end{aligned}$$

Deduce that $\sigma_\infty = 0.25|\text{Cloudy}\rangle + 0.5|\text{Sunny}\rangle + 0.25|\text{Rainy}\rangle$ and double-check that $t \gg\!= \sigma_\infty = \sigma_\infty$.

7.4.3 (The set-up of this exercise is copied from machine learning lecture notes of Doina Precup.) Consider a 5-state hallway of the form:



Thus we use a space $X = \{1, 2, 3, 4, 5\}$ of positions, together with a space $Y = \{2, 3\}$ of outputs, for the number of surrounding walls. The transition and emission channels $t \colon X \nrightarrow X$ and $e \colon X \nrightarrow Y$ for a robot in this hallway are given by:

$$\begin{aligned}t(1) &= \tfrac{3}{4}|1\rangle + \tfrac{1}{4}|2\rangle & e(1) &= 1|3\rangle \\ t(2) &= \tfrac{1}{4}|1\rangle + \tfrac{1}{2}|2\rangle + \tfrac{1}{4}|3\rangle & e(2) &= 1|2\rangle \\ t(3) &= \tfrac{1}{4}|2\rangle + \tfrac{1}{2}|3\rangle + \tfrac{1}{4}|4\rangle & e(3) &= 1|2\rangle \\ t(4) &= \tfrac{1}{4}|3\rangle + \tfrac{1}{2}|3\rangle + \tfrac{1}{4}|5\rangle & e(4) &= 1|2\rangle \\ t(5) &= \tfrac{1}{4}|4\rangle + \tfrac{3}{4}|5\rangle & e(5) &= 1|3\rangle.\end{aligned}$$

We use $\sigma = 1|3\rangle$ as start state, and we have a sequence of observations $\alpha = [2, 2, 3, 2, 3, 3]$, formally as a sequence of point predicates $[\mathbf{1}_2, \mathbf{1}_2, \mathbf{1}_3, \mathbf{1}_2, \mathbf{1}_3, \mathbf{1}_3]$.

1 Check that $(\sigma, t, e) \models \alpha = \frac{3}{512}$.

2 Next we filter with the sequence α. Show that it leads succesively to the following states σ_i as in Definition 7.4.4

$$\begin{aligned}
\sigma_1 \coloneqq \sigma &= 1|3\rangle \\
\sigma_2 &= \tfrac{1}{4}|2\rangle + \tfrac{1}{2}|3\rangle + \tfrac{1}{4}|4\rangle \\
\sigma_3 &= \tfrac{1}{16}|1\rangle + \tfrac{1}{4}|2\rangle + \tfrac{3}{8}|3\rangle + \tfrac{1}{4}|4\rangle + \tfrac{1}{16}|5\rangle \\
\sigma_4 &= \tfrac{3}{8}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|4\rangle + \tfrac{3}{8}|5\rangle \\
\sigma_5 &= \tfrac{1}{8}|1\rangle + \tfrac{1}{4}|2\rangle + \tfrac{1}{4}|3\rangle + \tfrac{1}{4}|4\rangle + \tfrac{1}{8}|5\rangle \\
\sigma_6 &= \tfrac{3}{8}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|4\rangle + \tfrac{3}{8}|5\rangle \\
\sigma_7 &= \tfrac{3}{8}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|4\rangle + \tfrac{3}{8}|5\rangle.
\end{aligned}$$

3 Show that a most likely sequence of states giving rise to the sequence of observations α is $[3, 2, 1, 2, 1, 1]$. Is there an alternative?

7.4.4 Apply Bayes' rule to the validity formulation (7.13) in order to prove the correctness of the following HMM validity algorithm.

$$\begin{aligned}
(\sigma, t, e) \models [] &\coloneqq 1 \\
(\sigma, t, e) \models [p] +\!\!+ \vec{p} &\coloneqq \Big(\sigma \models e \mathrel{=\!\!\ll} p\Big) \cdot \Big((t \gg\!= (\sigma|_{e \mathrel{=\!\!\ll} p}), t, e) \models \vec{p}\Big).
\end{aligned}$$

(Notice the connection with filtering from Definition 7.4.4.)

7.4.5 A *random walk* is a Markov model $d \colon \mathbb{Z} \rightsquigarrow \mathbb{Z}$ given by $d(n) = r|n-1\rangle + (1-r)|n+1\rangle$ for some $r \in [0, 1]$. This captures the idea that a step-to-the-left or a step-to-the-right are the only possible transitions. (The letter 'd' hints at modeling a drunkard.)

1 Start from initial state $\sigma = 1|0\rangle \in \mathcal{D}(\mathbb{Z})$ and describe a couple of subsequent states $d \gg\!= \sigma$, $d^2 \gg\!= \sigma$, $d^3 \gg\!= \sigma$, ... Which pattern emerges?

2 Prove that for $K \in \mathbb{N}$,

$$\begin{aligned}
d^K \gg\!= \sigma &= \sum_{0 \le k \le K} bn[K](1-r)(k) \Big| 2k - K \Big\rangle \\
&= \sum_{0 \le k \le K} \binom{K}{k} (1-r)^k r^{K-k} \Big| 2k - K \Big\rangle
\end{aligned}$$

7.4.6 A Markov chain $X \rightsquigarrow X$ has a 'one-stage history' only, in the sense that the state at stage $n+1$ depends only on the state at stage n. The following situation from [144, Chap. III, Ex. 4.4] involves a two-stage history.

Suppose that whether or not it rains today depends on wheather conditions through the last two days. Specifically, suppose that if it has rained for the past two days, then it will rain tomorrow with probability 0.7; if it rained today but not yesterday, then it will rain tomorrow with probability 0.5; if

it rained yesterday but not today, then it will rain tomorrow with probability 0.4; if it has not rained in the past to days, then it will rain tomorrow with probability 0.2.

1 Write $R = \{r, r^\perp\}$ for the state space of rain and no-rain outcomes, and capture the above probabilities via a channel $c \colon R \times R \nrightarrow R$.

2 Turn this channel c into a Markov chain $\langle \pi_2, c \rangle \colon R \times R \nrightarrow R \times R$, where the second component of $R \times R$ describes whether or not it rains on the current day, and the first component on the previous day. Describe $\langle \pi_2, c \rangle$ both as a function and as a string diagram.

3 Generalise this approach to a history of length $N > 1$: turn a channel $X^N \nrightarrow X$ into a Markov model $X^N \nrightarrow X^N$, where the relevant history is incorporated into the sample space.

7.4.7 Use the approach of the prevous exercise to turn a hidden Markov model into a Markov model.

7.4.8 In Definition 7.4.4 we have seen filtering for hidden Markov models, starting from a sequence of factors $\vec{p}$ as observations. In practice these factors are often point predicates $\mathbf{1}_{y_i}$ for a sequence of elements $\vec{y}$ of the visible space Y. Show that in that case one can describe filtering via Bayesian inversion as follows, for a hidden Markov model with transition channel $t \colon X \nrightarrow X$, emission channel $e \colon X \nrightarrow Y$ and initial state $\sigma \in \mathcal{D}(X)$.

$$\sigma_1 = \sigma \qquad \text{and} \qquad \sigma_{i+1} = (t \mathbin{\circ} e^\dagger_{\sigma_i})(y_i).$$

7.4.9 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two HMMs. Define their parallel product $\mathcal{H}_1 \otimes \mathcal{H}_2$ using the tensor operation $\otimes$ on states and channels.

7.5 Disintegration

In Subsections 1.3.3 and 1.4.3 we have seen how a binary relation $R \in \mathcal{P}(A \times B)$ on $A \times B$ corresponds to a $\mathcal{P}$-channel $A \to \mathcal{P}(B)$, and similarly how a multiset $\psi \in \mathcal{M}(A \times B)$ corresponds to an $\mathcal{M}$-channel $A \to \mathcal{M}(B)$. This phenomenon was called: extraction of a channel from a joint state. This section describes the analogue for probabilistic binary/joint states and channels (like in [52]). It turns out to be more subtle because probabilistic extraction requires normalisation — in order to ensure the unitality requirement of a $\mathcal{D}$-channel: multiplicities must add up to one.

In the probabilistic case this extraction is also called *disintegration*. It corresponds to a well known phenomenon, namely that one can write a joint probability $P(x, y)$ in terms of a conditional probability as via what is often called

the *chain rule*:

$$P(x, y) \;=\; P(y \mid x) \cdot P(x).$$

The mapping $x \mapsto P(y \mid x)$ is the channel involved, and $P(x)$ refers to the first marginal of $P(x, y)$. We have already seen this extraction of a channel via daggers, in Theorem 6.6.5.

Our current description of disintegration (and daggers) makes systematic use of the graphical calculus that we introduced earlier in this chapter. The formalisation of disintegration in category theory is not there (yet), but the closely related concept of Bayesian inversion — the dagger of a channel — has a nice categorical description, see Section 7.8 below.

We start by formulating the property of 'full support' in string diagrammatic terms. It is needed as pre-condition in disintegration.

Definition 7.5.1. Let X be a non-empty finite set. We say that a distribution or multiset ω on X has *full support* if $\omega(x) > 0$ for each $x \in X$.

More generally, we say that a $\mathcal{D}$- or $\mathcal{M}$-channel $c \colon Y \nrightarrow X$ has full support if the distribution/multiset $c(y)$ has full support for each y. This thus means that $c(y)(x) > 0$ for all $y \in Y$ and $x \in X$.

Lemma 7.5.2. *Let $\omega \in \mathcal{D}(X)$, where X is non-empty and finite. The following three points are equivalent:*

1. *ω has full support;*
2. *there is a predicate p on X and a non-zero probability $r \in (0, 1]$ such that $\omega(x) \cdot p(x) = r$ for each $x \in X$.*
3. *there is a predicate p on X such that $\omega|_p$ is the uniform distribution on X.*

Proof. Suppose that the finite set X has $N \geq 1$ elements. For the implication (1) $\Rightarrow$ (2), let $\omega \in \mathcal{D}(X)$ have full support, so that $\omega(x) > 0$ for each $x \in X$. Since $\omega(x) \leq 1$, we get $\frac{1}{\omega(x)} \geq 1$ so that $t \coloneqq \sum_x \frac{1}{\omega(x)} \geq N$ and $t \cdot \omega(x) \geq 1$, for each $x \in X$. Take $r = \frac{1}{t} \leq \frac{1}{N} \leq 1$, and put $p(x) = \frac{1}{t \cdot \omega(x)}$. Then:

$$\omega(x) \cdot p(x) \;=\; \omega(x) \cdot \tfrac{1}{t \cdot \omega(x)} \;=\; \tfrac{1}{t} \;=\; r.$$

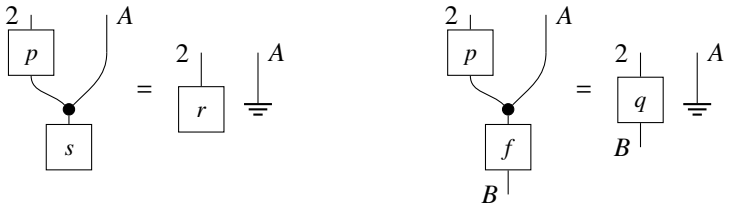
Next, for (2) $\Rightarrow$ (3), suppose $\omega(x) \cdot p(x) = r$, for some predicate p and non-zero probability r. Then $\omega \models p = \sum_x \omega(x) \cdot p(x) = N \cdot r$. Hence:

$$\omega|_p(x) \;=\; \frac{\omega(x) \cdot p(x)}{\omega \models p} \;=\; \frac{r}{N \cdot r} \;=\; \frac{1}{N} \;=\; \mathit{unif}_X(x).$$

For the final step (3) $\Rightarrow$ (1), let $\omega|_p = \mathit{unif}_X$. This requires that $\omega \models p \neq 0$ and thus that $\omega(x) \cdot p(x) = \frac{\omega \models p}{N} > 0$ for each $x \in X$. But then $\omega(x) > 0$. ❑

This result allows us to express the full support property in diagrammatic terms. This will be needed as precondition later on.

Definition 7.5.3. A state box s has *full support* if there is a predicate box p and a scalar box r, for some $r \in [0, 1]$, with an equation as on the left below.



Similarly, we say that a channel box f has *full support* if there are predicates p, q with an equation as above on the right.

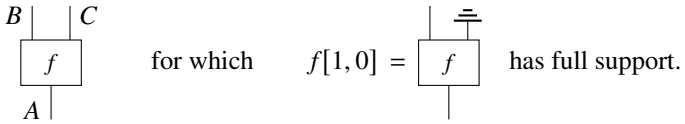
Strictly speaking, we didn't have to distinguish states and channels in this definition, since that state-case is a special instance of the channel-case, namely for $B = 1$. In the sequel we shall no longer make this distinction.

We now come to the main definition of this section. It involves turing a conditional probability $P(b, c \mid a)$ into $P(c \mid b, a)$ such that

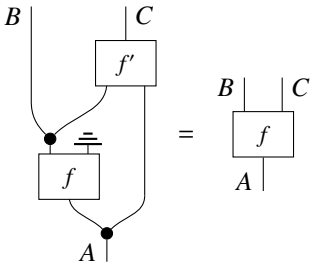
$$P(b, c \mid a) \;=\; P(c \mid b, a) \cdot P(b \mid a).$$

This is captured diagrammatically in (7.16) below.

Definition 7.5.4. Consider a box f



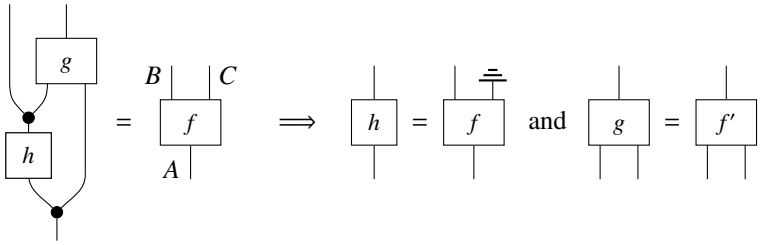
We say that f admits *disintegration* if there is a unique 'disintegration' box f' from B, A to C satisfying the equation:



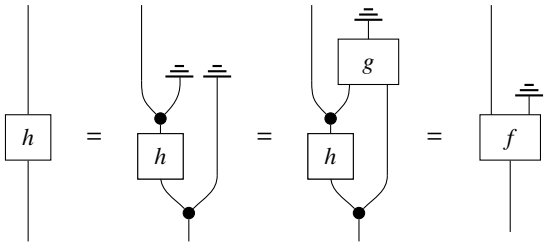
(7.16)

Uniqueness of f' in this definition means: for each box g from B, A to C and

for each box h from C to A with full support one has:



The first equation $h = f[1,0]$ in the conclusion is obtained simply by applying discard $\top$ to two right-most outgoing wires in the assumption — on the left and on the right of the equation — and using that g is unital:



The second equation $g = f'$ in the conclusion expresses uniqueness of the disintegration box f'.

We shall use the notation $dis_1(f)$ or $f[0,1\,|\,1,0]$ for this disintegration box f'. This notation will be explained in greater generality below.

The first thing to do is to check that this newly introduced construction is sound.

Proposition 7.5.5. *Disintegration as described in Definition 7.5.4 exists for probabilistic channels.*

Proof. Let $f\colon A \nrightarrow B \times C$ be a channel such that $f[1,0]\colon A \nrightarrow B$ has full support. The latter means that for each $a \in A$ and $b \in B$ one has $f[1,0](a)(b) = \sum_{c\in C} f(a)(b,c) \neq 0$. Then we can define $f'\colon B \times A \nrightarrow C$ as:

$$f'(b,a) \coloneqq \sum_c \frac{f(a)(b,c)}{f[1,0](a)(b)}\,|c\rangle. \tag{7.17}$$

By the full support assumption this is well-defined.

The left-hand side of Equation (7.16) evaluates as:

$$\begin{aligned}&\Big((id \otimes f') \circ (\Delta \otimes id) \circ (f[1,0] \otimes id) \circ \Delta\Big)(a)(b,c)\\ &= \textstyle\sum_{x,y} id(y)(b) \cdot f'(y,x)(c) \cdot f[1,0](a)(y) \cdot id(x)(a)\\ &= f'(b,a)(c) \cdot f[1,0](a)(b)\\ &= f(a)(b,c).\end{aligned}$$

For uniqueness, assume $(id \otimes g) \circ (\Delta \otimes id) \circ (h \otimes id) \circ \Delta = f$. As already mentioned, we can obtain $h = f[1,0]$ via diagrammatic reasoning. Here we reason element-wise. By assumption, $g(b,a)(c) \cdot h(a)(b) = f(a)(b,c)$. This gives:

$$\begin{aligned}h(a)(b) \;=\; 1 \cdot h(a)(b) \;=\; \big(\textstyle\sum_c g(b,a)(c)\big) \cdot h(a)(b) \;&=\; \textstyle\sum_c g(b,a)(c) \cdot h(a)(b)\\ &= \textstyle\sum_c f(a)(b,c)\\ &= f[1,0](a)(b).\end{aligned}$$

Hence $h = f[1,0]$. But then $g = f'$ since:

$$g(b,a)(c) \;=\; \frac{f(a)(b,c)}{h(a)(b)} \;=\; \frac{f(a)(b,c)}{f[1,0](a)(b)} \;=\; f'(b,a)(c). \qquad \blacksquare$$

Without the full support requirement, disintegrations may still exist, but they are not unique, see Example 7.6.1 (2) for an illustration.

Next we elaborate on the notation that we will use for disintegration.

Remark 7.5.6. In traditional notation in probability theory one simply omits variables to express marginalisation. For instance, for a distribution $\omega \in \mathcal{D}(X_1 \times X_2 \times X_3 \times X_4)$, considered as function $\omega(x_1, x_2, x_3, x_4)$ in four variables x_i, one writes:

$$\omega(x_2, x_3) \qquad \text{for the marginal} \qquad \sum_{x_1, x_4} \omega(x_1, x_2, x_3, x_4).$$

We have been using *masks* instead: lists containing as only elements 0 or 1. We then write $\omega[0,1,1,0] \in \mathcal{D}(X_2 \times X_3)$ to express the above marginalisation, with a 0 (resp. a 1) at position i in the mask meaning that the i-th variable in the distribution is discarded (resp. kept).

We like to use a similar mask-style notation for disintegration, mimicking the traditional notation $\omega(x_1, x_4 \mid x_2, x_3)$. This requires *two* masks, separated by the sign '|' for conditional probability. We sketch how it works for a box f from A to $B_1, \ldots, B_n$ in a string diagram. The notation

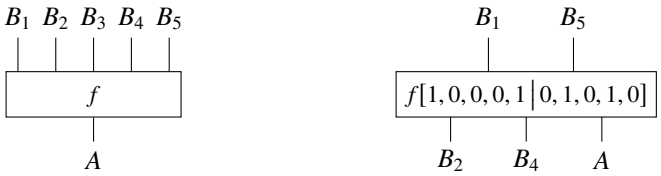
$$f[N \mid M]$$

will be used for a disintegration channel, in the following manner:

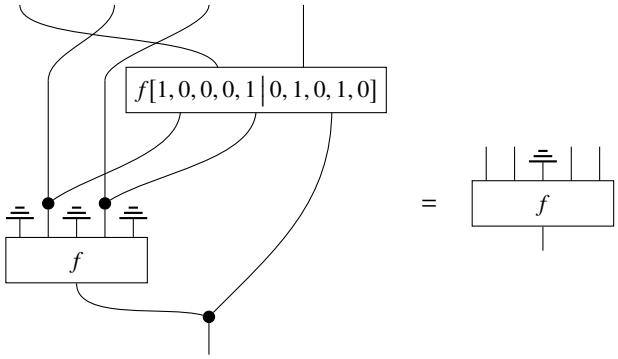
1 masks M, N must both be of length n;

2 M, N must be *disjoint*: there is no position i with a 1 both in M and in N;
3 the marginal $f[M]$ must have full support;
4 the domain of the disintegration box $f[N \mid M]$ is $\vec{B} \cap M, A$, where $\vec{B} \cap M$ contain precisely those B_i with a 1 in M at position i;
5 the codomain of $f[N \mid M]$ is $\vec{B} \cap N$.
6 $f[N \mid M]$ is unique in satisfying an "obvious" adaptation of Equation (7.16), in which $f[M \cup N]$ is equal to a string diagram consisting of $f[M]$ suitably followed by $f[N \mid M]$.

How this really works is best illustrated via a concrete example. Consider the box f on the left below, with five output wires. We elaborate the disintegration $f[1, 0, 0, 0, 1 \,|\, 0, 1, 0, 1, 0]$, as on the right.



The disintegration box $f[1, 0, 0, 0, 1 \,|\, 0, 1, 0, 1, 0]$ is unique in satisfying:



In traditional notation one could express this equation as:

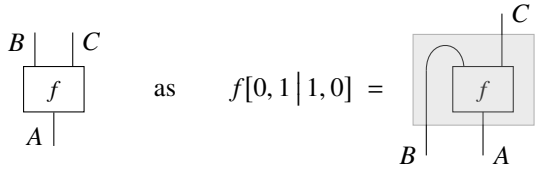
$$f(b_1, b_5 \mid b_2, b_4, a) \cdot f(b_2, b_4 \mid a) \;=\; f(b_1, b_2, b_4, b_5 \mid a).$$

Two points are still worth noticing.

- It is not required that at position i there is a 1 either in mask M or in mask N when we form $f[N \mid M]$. If there is a 0 at i both in M and in N, then the wire at i is discarded altogether. This happens in the above illustration for the third wire: the string diagram on the right-hand side of the equation is $f[M \cup N] = f[1, 1, 0, 1, 1]$.

- The above disintegration $f[1,0,0,0,1 \,|\, 0,1,0,1,0]$ can be obtained from the 'simple', one-wire version of disintegration in Definition 7.5.4 by first suitably rearranging wires and combining them via products. How to do this precisely is left as an exercise (see below).

Disintegration is a not a 'compositional' operation that can be obtained by combining other string diagrammatic primitives. The reason is that disintegration involves normalisation, via division in (7.17). Still it would be nice to be able to use disintegration in diagrammatic form. For this purpose one may use a trick: in the setting of Definition 7.5.4 we 'bend' the relevant wire downwards and put a gray box around the result in order to suggest that its interior is closed off and has become inaccessible. Thus, we write the disintegration of:



This 'shaded box' notation can also be used for more complicated forms of disintegration, as described above. This notation is useful to express some basic properties of disintegration, see Exercise 7.5.4.

Exercises

7.5.1 1 Prove that if a joint state has full support, then each of its marginals has full support.

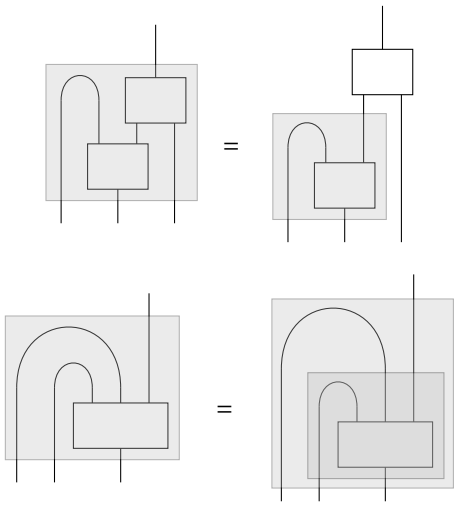
2 Consider the joint state $\omega = \frac{1}{2}|a,b\rangle + \frac{1}{2}|a^\perp,b^\perp\rangle \in \mathcal{D}(A \times B)$ for $A = \{a,a^\perp\}$ and $B = \{b,b^\perp\}$. Check that both marginals $\omega[1,0] \in \mathcal{D}(A)$ and $\omega[0,1] \in \mathcal{D}(B)$ have full support, but ω itself does not. Conclude that the converse of the previous point does not hold.

7.5.2 Consider the box f in Remark 7.5.6. Write down the equation for the disintegration $f[0,1,0,1,0 \,|\, 1,0,1,0,1]$. Formulate also what uniqueness means.

7.5.3 Show how to obtain the disintegration $f[0,0,0,1,1 \,|\, 0,1,1,0,0]$ in Remark 7.5.6 from the formulation in Definition 7.5.4 via rearranging and combining wires (via $\times$).

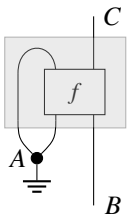
7.5.4 (From [52]) Prove the following 'sequential' and 'parallel' properties of disintegration. The best is to give a diagrammatic proof, using uniqueness of disintegration. But you may also check that these equations are sound, *i.e.* hold for all interpretations of the boxes as

probabilistic channels (with appropriate full support).



7.5.5 Let $f\colon A \times B \to A \times C$ be a channel, where A, B, C are finite sets. Define the "probabilistic trace" $prtr(f)\colon B \to C$ as:

$$prtr(f) \quad := \tag{7.18}$$



1 Check that:

$$prtr(f)(b)(c) = \sum_a \frac{1}{\#A} \cdot \frac{f(a,b)(a,c)}{\sum_y f(a,b)(a,y)},$$

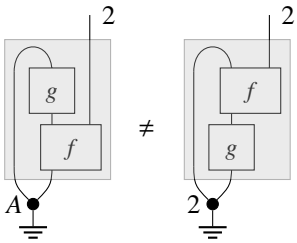
where $\#A$ is the number of elements of A.
A categorically oriented reader might now ask if the above definition $prtr(f)$ yields a proper 'trace' construction in the symmetric monoidal Kleisli category $\mathbf{Chan}(\mathcal{D})$ of probabilistic channels, but this is not the case: the so-called dinaturality condition fails. This will be illustrated in the next few points.

2 Take space $A = \{a, b, c\}$ and $2 = \{0, 1\}$ with channels $f\colon A \to 2 \times 2$ and $g\colon 2 \to A$ defined by:

$$\begin{aligned} f(a) &= \tfrac{1}{4}|0,0\rangle + \tfrac{3}{4}|1,0\rangle \\ f(b) &= \tfrac{2}{5}|0,0\rangle + \tfrac{3}{5}|1,1\rangle \\ f(c) &= \tfrac{1}{2}|0,1\rangle + \tfrac{1}{2}|1,0\rangle \end{aligned} \qquad \begin{aligned} g(0) &= \tfrac{1}{3}|a\rangle + \tfrac{2}{3}|c\rangle \\ g(1) &= \tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle. \end{aligned}$$

The aim is to show an inequality of states:

$$prtr((g \otimes id) \circ f) \;=\; \;\neq\; \;=\; prtr(f \circ g)$$



Both sides are an instance of (7.18) with $B = 1$.

3 Check that $(g \otimes id) \circ f \colon A \nrightarrow A \times 2$ is:

$$\begin{aligned} a &\mapsto \tfrac{11}{24}|a,0\rangle + \tfrac{3}{8}|b,0\rangle + \tfrac{1}{6}|c,0\rangle \\ b &\mapsto \tfrac{2}{15}|a,0\rangle + \tfrac{3}{10}|a,1\rangle + \tfrac{3}{10}|b,1\rangle + \tfrac{4}{15}|c,0\rangle \\ c &\mapsto \tfrac{1}{4}|a,0\rangle + \tfrac{1}{6}|a,1\rangle + \tfrac{1}{4}|b,0\rangle + \tfrac{1}{3}|c,1\rangle. \end{aligned}$$

And that $f \circ g \colon 2 \nrightarrow 2 \times 2$ is:

$$\begin{aligned} 0 &\mapsto \tfrac{1}{12}|0,0\rangle + \tfrac{1}{3}|0,1\rangle + \tfrac{7}{12}|1,0\rangle \\ 1 &\mapsto \tfrac{13}{40}|0,0\rangle + \tfrac{3}{8}|1,0\rangle + \tfrac{3}{10}|1,1\rangle. \end{aligned}$$

4 Now show that the disintegration $((g \otimes id) \circ f)[0,1 \,|\, 1,0] \colon A \times A \nrightarrow 2$ is:

$$\begin{array}{lll} (a,a) \mapsto 1|0\rangle & (b,a) \mapsto \tfrac{4}{13}|0\rangle + \tfrac{9}{13}|1\rangle & (c,a) \mapsto \tfrac{3}{5}|0\rangle + \tfrac{2}{5}|1\rangle \\ (a,b) \mapsto 1|0\rangle & (b,b) \mapsto 1|1\rangle & (c,b) \mapsto 1|0\rangle \\ (a,c) \mapsto 1|0\rangle & (b,c) \mapsto 1|0\rangle & (c,c) \mapsto 1|1\rangle. \end{array}$$

And that $(f \circ g)[0,1 \,|\, 1,0] \colon 2 \times 2 \nrightarrow 2$ is:

$$\begin{array}{ll} (0,0) \mapsto \tfrac{1}{5}|0\rangle + \tfrac{4}{5}|1\rangle & (1,0) \mapsto 1|0\rangle \\ (0,1) \mapsto 1|0\rangle & (1,1) \mapsto \tfrac{5}{9}|0\rangle + \tfrac{4}{9}|1\rangle. \end{array}$$

5 Conclude that:

$$prtr((g \otimes id) \circ f) \;=\; \tfrac{1}{3}|0\rangle + \tfrac{2}{3}|1\rangle,$$

whereas:

$$prtr(f \circ g) \;=\; \tfrac{17}{45}|0\rangle + \tfrac{28}{45}|1\rangle.$$

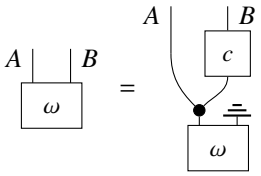
7.6 Disintegration for states

The previous section has described disintegration for channels, but in the literature it is mostly used for states only, that is, for boxes without incoming wires. This section will look at this special case, following [21]. It will reformulate the channel extraction from Theorem 6.6.5 via string diagrams.

In its most basic form disintegration involves extracting a channel from a (binary) joint state, so that the state can be reconstructed as a graph — as in Definition 2.5.6. This works as follows.



so that:



(7.19)

Using the notation for disintegration from Remark 7.5.6 we can write the extracted channel c as $c = \omega[0,1 \,|\, 1,0]$. Occasionaly we also write it as $c = dis_1(\omega)$. We can express Equation (7.19) as graph equation

$$\begin{aligned} \omega &= (id \otimes \omega[0,1 \,|\, 1,0]) \gg\!= (\Delta \gg\!= \omega[1,0]) \\ &= \langle id,\, \omega[0,1 \,|\, 1,0]\rangle \gg\!= \omega[1,0], \end{aligned} \tag{7.20}$$

like in Theorem 6.6.5. This is a special case of Equation (7.16), where the incoming wire is trivial (equal to 1). We should not forget the side-condition of disintegration, which in this case is: the first marginal $\omega[1,0]$ must have full support. Then one can define the channel $c = dis_1(\omega) = \omega[0,1 \,|\, 1,0] \colon A \nrightarrow B$ as disintegration, explicitly, as a special case of (7.17):

$$c(a) \coloneqq \sum_{b \in B} \frac{\omega(a,b)}{\sum_y \omega(a,y)} \big|b\big\rangle = \sum_{b \in B} \frac{\omega(a,b)}{\omega[1,0](a)} \big|b\big\rangle. \tag{7.21}$$

From a joint state $\omega \in \mathcal{D}(X \times Y)$ we can extract a channel $\omega[0,1 \,|\, 1,0] \colon X \nrightarrow Y$ and also a channel $\omega[1,0 \,|\, 0,1] \colon Y \nrightarrow X$ in the opposite direction, provided that ω's marginals have full support, see again Theorem 6.6.5. The direction of the channels is thus in a certain sense arbitrary, and does not reflect any form of causality, see Chapter **??**.

Example 7.6.1. We shall look at two examples, involving spaces $A = \{a, a^\perp\}$ and $B = \{b, b^\perp\}$.

1 Consider the following state $\omega \in \mathcal{D}(A \times B)$,

$$\omega = \tfrac{1}{4}|a, b\rangle + \tfrac{1}{2}|a, b^\perp\rangle + \tfrac{1}{4}|a^\perp, b^\perp\rangle.$$

We have as first marginal $\sigma := \omega[1, 0] = \frac{3}{4}|a\rangle + \frac{1}{4}|a^\perp\rangle$ with full support. The extracted channel $c := \omega[0, 1 \,|\, 1, 0] \colon A \nrightarrow B$ is given by:

$$c(a) = \frac{\omega(a,b)}{\sigma(a)}|b\rangle + \frac{\omega(a,b^\perp)}{\sigma(a)}|b^\perp\rangle = \frac{1/4}{3/4}|b\rangle + \frac{1/2}{3/4}|b^\perp\rangle = \tfrac{1}{3}|b\rangle + \tfrac{2}{3}|b^\perp\rangle$$
$$c(a^\perp) = \frac{\omega(a^\perp,b)}{\sigma(a^\perp)}|b\rangle + \frac{\omega(a^\perp,b^\perp)}{\sigma(a^\perp)}|b^\perp\rangle = \frac{0}{1/4}|b\rangle + \frac{1/4}{1/4}|b^\perp\rangle = 1|b^\perp\rangle.$$

Then indeed, $\langle id, c\rangle \gg\!= \sigma = \omega$.

2 Now let's start from:

$$\omega = \tfrac{1}{3}|a, b\rangle + \tfrac{2}{3}|a, b^\perp\rangle.$$

Then $\sigma := \omega[1, 0] = 1|a\rangle$. It does not have full support. Let $\tau \in \mathcal{D}(B)$ be an arbitrary state. We define $c \colon A \nrightarrow B$ as:

$$c(a) = \tfrac{1}{3}|b\rangle + \tfrac{2}{3}|b^\perp\rangle \qquad \text{and} \qquad c(a^\perp) = \tau.$$

We then still get $\langle id, c\rangle \gg\!= \sigma = \omega$, no matter what τ is.

More generally, if we don't have full support, disintegrations still exist, but they are not unique. We generally avoid such non-uniqueness by requiring full support.

Example 7.6.2. The *natural join* $\bowtie$ is a basic construction in database theory that makes it possible to join two databases which coincide on their overlap. Such natural joins are used in a probabilistic setting in [1, 16] — in particular in relation to Bell tables, see Exercises 2.2.11, **??**.

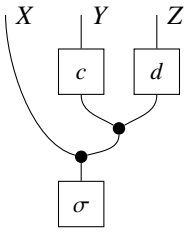
Let $\omega \in \mathcal{D}(X \otimes Y)$ and $\rho \in \mathcal{D}(X \otimes Z)$ be two joint states with equal X-marginal: $\omega[1, 0] = \rho[1, 0]$. A *natural join*, if it exists, is a state $\omega \bowtie \rho \in \mathcal{D}(X \otimes Y \otimes Z)$ which marginalises both to ω and two ρ, as in:

$$(\omega \bowtie \rho)[1, 1, 0] = \omega \qquad \text{and} \qquad (\omega \bowtie \rho)[1, 0, 1] = \rho.$$

We show how such natural joins can be constructed via disintegration of states. Let's assume we have joint states ω and ρ as described above, with common marginal written as $\sigma := \omega[1, 0] = \rho[1, 0]$. We extract channels:

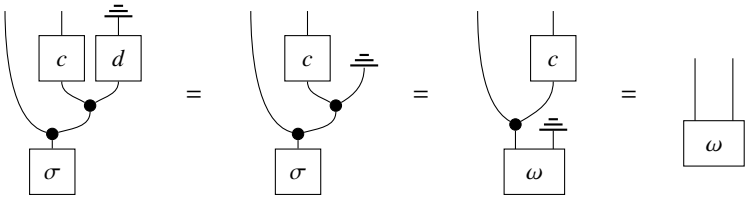
$$c := \omega[0, 1 \,|\, 1, 0] : X \nrightarrow Y \qquad d := \rho[0, 1 \,|\, 1, 0] : X \nrightarrow Z.$$

Now we define:

$\omega \bowtie \rho \quad :=$ 

(7.22)

The $(\omega \bowtie \rho)[1, 1, 0] = \omega$ can be obtained easily via diagrammatic reasoning:



Similarly one proves $(\omega \bowtie \rho)[1, 0, 1] = \rho$. For a concrete instantiation of this construction, see Exercise 7.6.6. Such natural joins are typically non-trivial, but the above construction (7.22) gives a clear recipe. The graphical approach clearly describes the relevant flows and is especially useful in more complicated situations, with multiple states which agree on multiple marginals.

The next result illustrates two bijective correspondences resulting from disintegration. These correspondences will be indicated with a double line $\;=\!=\!=\;$, like in (1.3), expressing that one can translate in two directions, from what is above the lines to what is below, and vice-versa. The first bijective correspondence is basically a reformulation of disintegration itself (for states). The second one is new and involves distributions over distributions — sometimes called hyperdistributions [121, 122].

Theorem 7.6.3. *Let X, Y be arbitrary sets, where Y is finite.*

1 There is a bijective correspondence between:

$$\frac{\tau \in \mathcal{D}(X \times Y) \text{ where } \tau[0, 1] \text{ has full support}}{\omega \in \mathcal{D}(X) \text{ and } c \colon X \to Y \text{ such that } c \gg\!= \omega \text{ has full support}}$$

2 For each natural number $N \geq 1$ there is a bijective correspondence between:

$$\frac{\Omega \in \mathcal{D}(\mathcal{D}(X)) \text{ with } \big|\mathrm{supp}(\Omega)\big| = N}{\rho \in \mathcal{D}(X \times N) \text{ where } \rho[0, 1] \text{ has full support}}$$

Recall that we write $|A| \in \mathbb{N}$ for the number of elements of a finite set A.

Proof. 1 In the downward direction, starting from a joint state $\tau \in \mathcal{D}(X \times Y)$ we obtain a state $\omega \coloneqq \tau[1,0] \in \mathcal{D}(X)$ as first marginal and a channel $c \coloneqq \tau[0,1 \,|\, 1,0] \colon X \nrightarrow Y$ by disintegration. The latter exists because the second marginal $\tau[0,1]$ has full support. In the upward direction we transform a state-channel pair ω, c to the joint state $\tau \coloneqq \langle id, c \rangle \gg\!= \omega \in \mathcal{D}(X \times Y)$. By assumption, $\tau[0,1] = c \gg\!= \omega$ has full support. Doing these transformations twice, both down-up and up-down, yields the orginal data, by definition of disintegration.

2 Let distribution $\Omega \in \mathcal{D}(\mathcal{D}(X))$ have a support with N elements, say $supp(\Omega) = \{\omega_0, \ldots, \omega_{N-1}\}$ with $\omega_i \in \mathcal{D}(X)$. We take:

$$\rho(x,i) \coloneqq \Omega(\omega_i) \cdot \omega_i(x). \tag{7.23}$$

This yields a distribution $\rho \in \mathcal{D}(X \times N)$ since these probabilities add up to one:

$$\sum_{x \in X,\, i \in N} \rho(x,i) \coloneqq \sum_{i \in N} \Omega(\omega_i) \cdot \sum_{x \in X} \omega_i(x) = \sum_{i \in N} \Omega(\omega_i) = 1.$$

Furthermore, ρ's second marginal has full support, since for each $i \in N$,

$$\rho[0,1](i) = \sum_{x \in X} \rho(x,i) = \Omega(\omega_i) \cdot \sum_{x \in X} \omega_i(x) = \Omega(\omega_i) > 0.$$

In the upward direction, starting from $\rho \in \mathcal{D}(X \times N)$ we from the channel $d \coloneqq \rho[1,0 \,|\, 0,1] \colon N \to X$ by disintegration and use it to define $\Omega \in \mathcal{D}(\mathcal{D}(X))$ as:

$$\Omega \coloneqq \sum_{i \in N} \rho[0,1](i) \,\big|\, d(i) \big\rangle. \tag{7.24}$$

Since $\rho[0,1](i) > 0$ for each i, this Ω's support has N elements.

Starting from Ω with support $\{\omega_0, \ldots, \omega_{N-1}\}$, we can get ρ as in (7.23) with extracted channel d satisfying:

$$d(i)(x) = \frac{\rho(x,i)}{\rho[0,1](i)} = \frac{\Omega(\omega_i) \cdot \omega_i(x)}{\Omega(\omega_i)} = \omega_i(x).$$

Hence the definition (7.24) yields the original state $\Omega \in \mathcal{D}(\mathcal{D}(X))$:

$$\sum_{i \in N} \rho[0,1](i) \big| d(i) \big\rangle = \sum_{i \in N} \Omega(\omega_i) \big| \omega_i \big\rangle = \Omega.$$

In the other direction, starting from a joint state $\rho \in \mathcal{D}(X \times N)$ whose second marginal has full support, we can form Ω as in (7.24) and turn it into a joint state again via (7.23), whose probability at (x,i) is:

$$\begin{aligned} \Omega(d(i)) \cdot d(i)(x) &= \rho[0,1](i) \cdot \rho[1,0 \,\big|\, 0,1](i)(x) \\ &= \Big(\langle id, \rho[1,0 \,\big|\, 0,1] \rangle \gg\!= \rho[0,1] \Big)(x,i) \\ &= \rho(x,i). \end{aligned}$$

This last equation holds by disintegration. ❑

By combining the two correspondences in this theorem we can bijectively relate a hyperdistribution $\Omega \in \mathcal{D}(\mathcal{D}(X))$ and a state-channel pair $\omega \in \mathcal{D}(X)$, $c \colon X \nrightarrow Y$, see Exercise 7.6.7 below. This corresponce is used in Bayesian persuasion, see [97], where the channel c is called a signal.

With disintegration well-understood we can formulate a follow-up of the crossover update corollary 6.3.13. There we looked at *marginalisation* after update. Here we look at *extraction* of a channel after such an update, at the same position of the original channel. The newly extracted channel can be expressed in terms of an updated channel, see Definition 6.1.1 (3).

Theorem 7.6.4. *Let $c \colon X \nrightarrow Y$ be a channel with state $\sigma \in \mathcal{D}(X)$ on its domain, and let $p \in \mathsf{Fact}(X)$ and $q \in \mathsf{Fact}(Y)$ be factors.*

1 The extraction on an updated graph state yields:

$$(\langle id, c\rangle \gg\!= \sigma)\Big|_{p\otimes q}[0, 1 \,\big|\, 1, 0] \;=\; c|_q \qquad \text{where} \qquad c|_q(x) \coloneqq c(x)|_q.$$

2 And as a result:

$$(\langle id, c\rangle \gg\!= \sigma)\Big|_{p\otimes q} \;=\; \langle id, c|_q\rangle \gg\!= \sigma|_{p \,\&\, (c \prec\!\ll q)}.$$

Proof. 1 For $x \in X$ and $y \in Y$ we have:

$$\begin{aligned}
(\langle id, c\rangle \gg\!= \sigma)\Big|_{p\otimes q}[0, 1 \,\big|\, 1, 0](x)(y) &\overset{(7.21)}{=} \frac{(\langle id, c\rangle \gg\!= \sigma)\Big|_{p\otimes q}(x, y)}{(\langle id, c\rangle \gg\!= \sigma)\Big|_{p\otimes q}[1, 0](x)} \\
&= \frac{(\langle id, c\rangle \gg\!= \sigma)(x, y) \cdot (p \otimes q)(x, y)}{\sum_v (\langle id, c\rangle \gg\!= \sigma)(x, v) \cdot (p \otimes q)(x, v)} \\
&= \frac{\sigma(x) \cdot c(x)(y) \cdot p(x) \cdot q(y)}{\sum_v \sigma(x) \cdot c(x)(v) \cdot p(x) \cdot q(v)} \\
&= \frac{c(x)(y) \cdot q(y)}{\sum_v c(x)(v) \cdot q(v)} \\
&= \frac{c(x)(y) \cdot q(y)}{c(x) \models q} \\
&= c(x)|_q(y) \\
&= c|_q(x)(y).
\end{aligned}$$

2 An (updated) joint state like $(\langle id, c\rangle \gg\!= \sigma)|_{p\otimes q}$ can always be written as graph $\langle id, d\rangle \gg\!= \tau$. The channel d is obtained by distintegration of the joint state, and equals $c|_q$ by the previous point. The state τ is the first marginal of the

joint state. Hence we are done by charactersing this first marginal, in:

$$\begin{aligned}
&(\langle id, c\rangle \gg\!= \sigma)\big|_{p\otimes q}[1,0] \\
&= (\langle id, c\rangle \gg\!= \sigma)\big|_{(\mathbf{1}\otimes q)\&(p\otimes \mathbf{1})}[1,0] && \text{by Exercise 4.3.7} \\
&= (\langle id, c\rangle \gg\!= \sigma)\big|_{\mathbf{1}\otimes q}\big|_{p\otimes \mathbf{1}}[1,0] && \text{by Lemma 6.1.6 (3)} \\
&= (\langle id, c\rangle \gg\!= \sigma)\big|_{\mathbf{1}\otimes q}[1,0]\big|_{p} && \text{by Lemma 6.1.6 (6)} \\
&= \sigma|_{c \prec\!\ll q}|_{p} && \text{by Corollary 6.3.13 (2)} \\
&= \sigma|_{p\,\&\,(c \prec\!\ll q)} && \text{by Lemma 6.1.6 (3).}
\end{aligned}$$ ❑

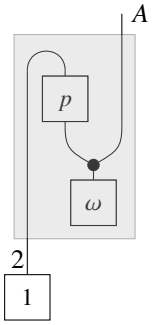
7.6.1 Disintegration of states and conditioning

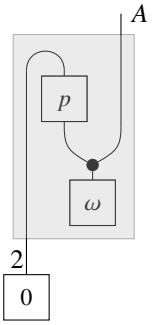
With disintegration (of states) we can express conditioning (updating) of states in a new way. Also, the earlier results on crossover inference can now be reformulated via conditioning.

For the next result, recall that we can identify a predicate $X \to [0,1]$ with a channel $X \to\!\!\!\!\bullet\; 2$, see Exercise 4.3.5.

Proposition 7.6.5. *Let $\omega \in \mathcal{D}(A)$ be state and $p \in \mathrm{Pred}(A)$ be predicate, both on the same finite set A. Assumme that the validity $\omega \models p$ is neither 0 nor 1.*

Then we can express the conditionings $\omega|_p \in \mathcal{D}(A)$ and $\omega|_{p^\perp} \in \mathcal{D}(A)$ as (interpreted) string diagrams:

$\omega|_p$ = 

and $\omega|_{p^\perp}$ = 

(7.25)

Proof. First, let's write $\sigma \coloneqq \langle p, id\rangle \gg\!= \omega \in \mathcal{D}(2 \times A)$ for the joint state that is disintegrated in the above string diagrams. The pre-conditioning for disintegration is that the marginal $\sigma[1,0] \in \mathcal{D}(2)$ has full support. Explicitly, this means for $b \in 2 = \{0,1\}$,

$$\sigma[1,0](b) = \sum_{a\in A} \sigma(b,a) = \sum_{a\in A} \omega(a)\cdot p(a)(b) = \begin{cases} \omega \models p & \text{if } b = 1 \\ \omega \models p^\perp & \text{if } b = 0. \end{cases}$$

The full support requirement that $\sigma[1,0](b) > 0$ for each $b \in 2$ means that both $\omega \models p$ and $\omega \models p^\perp = 1 - (\omega \models p)$ are non-zero. This holds by assumption.

We elaborate the above string diagram on the left. Let's write c for the extracted channel. It is, according to (7.21),

$$c(1) = \sum_{a \in A} \frac{\sigma(1,a)}{\sum_a \sigma(1,a)} \big|a\big\rangle = \sum_{a \in A} \frac{\omega(a) \cdot p(a)}{\omega \models p} \big|a\big\rangle = \omega|_p. \qquad \blacksquare$$

We move to crossover inference, as described in Corollary 6.3.13. There, a joint state $\omega \in \mathcal{D}(X \times Y)$ is used, which can be described as a graph $\omega = \langle id, c\rangle \gg\!= \sigma$. But we can drop this graph assumption now, since it can be obtained from ω via disintegration — assuming that ω's first marginal has full support.

Thus we come to the following reformulation of Corollary 6.3.13.

Theorem 7.6.6. *Let $\omega \in \mathcal{D}(X \times Y)$ be joint state whose first marginal $\sigma := \omega[1,0] \in \mathcal{D}(X)$ has full support, so that the channel $c := \omega[0,1 \,\big|\, 1,0] \colon X \nrightarrow Y$ exists, uniquely, by disintegration, and thus $\omega = \langle id, c\rangle \gg\!= \sigma$. Then:*

1 for a factor p on X,

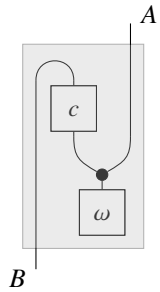
$$(\omega|_{p \otimes \mathbf{1}})[0,1] = c \gg\!= \sigma|_p.$$

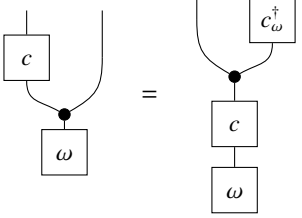
2 for a factor q on Y,

$$(\omega|_{\mathbf{1} \otimes q})[1,0] = \sigma|_{c \ll q}. \qquad \blacksquare$$

7.6.2 Bayesian inversion, graphically

In Section 6.6 we have introduced Bayesian inversion as the dagger of a channel. Here we redescribe it in graphical form. So let $c \colon A \nrightarrow B$ be a channel with finite codomain B, and with a state $\omega \in \mathcal{D}(A)$ on its domain, such that the predicted state $c \gg\!= \omega \in \mathcal{D}(B)$ has full support. The Bayesian inversion (also called: dagger) of c wrt. ω is the channel $c^\dagger_\omega \colon B \nrightarrow A$ defined as disintegration of the joint state $\langle c, id\rangle \gg\!= \omega$ in:

$c^\dagger_\omega :=$ 

such that 

(7.26)

By construction, $c^\dagger_\omega$ is the unique box giving the above equation. By applying

$\mathrel{\overline{+}}$ to the left outgoing line on both sides of the above equation we get the equation that we saw earlier in Proposition 6.6.4 (1):

$$\omega = c^{\dagger}_{\omega} \gg\!= (c \gg\!= \omega).$$

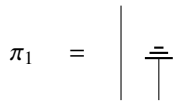
Above, we have described Bayesian inversion in terms of disintegration. In fact, the two notions are inter-definable, since disintegration can also be defined in terms of Bayesian inversion. This is in fact what we have done in Theorem 6.6.5. We now put it in graphical perspective.

Lemma 7.6.7. *Let $\omega \in \mathcal{D}(A \times B)$ be such that its first marginal $\pi_1 \gg\!= \omega = \omega[1, 0] \in \mathcal{D}(A)$ has full support, where $\pi_1 : A \times B \to A$ is the projection channel. The second marginal of the Bayesian inversion:*

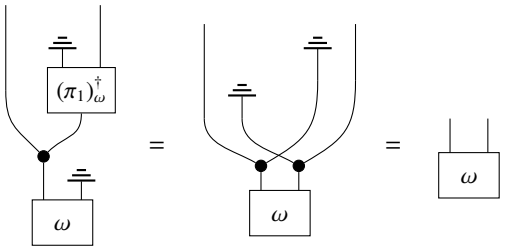
$$\pi_2 \gg\!= (\pi_1)^{\dagger}_{\omega} \qquad \text{for} \qquad (\pi_1)^{\dagger}_{\omega} : A \to A \times B,$$

is then the disintegration channel $A \to B$ for ω.

Proof. We first note that the projection channel $\pi_1 : A \times B \to A$ can be written as string diagram:



We need to prove the equation in (7.19). It can be obtained via purely diagrammatic reasoning:



The first equation is an instance of (7.26). ❑

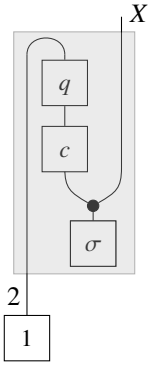
7.6.3 Pearl's and Jeffrey's update rules, graphically

In Section 6.7 we first described the update rules of Pearl and Jeffrey. At this stage we can recast them in terms of string diagrams. Recall that the setting of these rules is given by a channel $c : X \to Y$ with a state $\sigma \in \mathcal{D}(Y)$ on its domain. We are presented with evidence on Y and like to transport it along c in order to update σ.

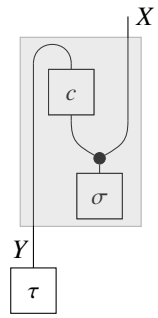
In Pearl's update rule the evidence is given by a factor. Here we restrict it to a predicate $q \colon Y \to 2$, since such predicates can be represented in a string diagram. The state σ is then updated to $\sigma|_{c \impliedby q}$. This is represented on the left below. It corresponds to the update diagram (on the left) in (7.25), but now with transformed predicate $c \impliedby q$.

For Jeffrey's update we have evidence in the form of a state $\tau \in \mathcal{D}(Y)$ and we update σ to $c^{\dagger}_{\sigma} \gg\!= \tau$. Hence we can use the string diagram of the dagger, as in (7.26). The result is on the right.

Pearl's update rule:



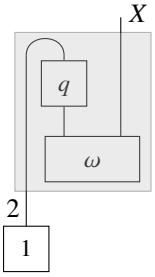
Jeffrey's update rule:



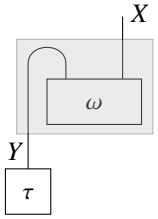
(7.27)

We see that the state σ and channel $c \colon X \to Y$ together form a graph state $\omega \coloneqq \langle c, id \rangle \gg\!= \sigma \in \mathcal{D}(Y \times X)$. This joint state ω is what's really used in the above two diagrams. Hence we can also reformulate the above to update diagrams in terms of this joint state ω.

Pearl:



Jeffrey:



(7.28)

This 'joint' formulation is sometimes convenient when the situation at hand is presented in terms of a joint state, like in Exercise 6.7.3.

Exercises

7.6.1 Check that Proposition 7.6.5 can be reformulated as:

$$\omega|_p = p^{\dagger}_{\omega}(1) \qquad \text{and} \qquad \omega|_{p^{\perp}} = p^{\dagger}_{\omega}(0).$$

7.6.2 Let $\sigma \in \mathcal{D}(X)$ have full support and consider $\omega := \sigma \otimes \tau$ for some $\tau \in \mathcal{D}(Y)$. Check that the channel $X \nrightarrow Y$ extracted from ω by disintegration is the constant function $x \mapsto \tau$. Give a string diagrammatic account of this situation.

7.6.3 Consider an extracted channel $\omega[1, 0, 0 \,|\, 0, 0, 1]$ for some state ω.

1 Write down the defining equation for this channel, as string diagram.

2 Check that $\omega[1, 0, 0 \,|\, 0, 0, 1]$ is the same as $\omega[1, 0, 1][1, 0 \,|\, 0, 1]$.

7.6.4 Check that a marginalisation ωM can also be described as disintegration $\omega[M \,|\, 0, \ldots, 0]$ where the number of 0's equals the length of the mask/list M.

7.6.5 Disintegrate the distribution $\mathit{Flrn}(\tau) \in \mathcal{D}(\{H, L\} \times \{1, 2, 3\})$ in Subsection 1.4.1 to a channel $\{H, L\} \nrightarrow \{1, 2, 3\}$.

7.6.6 Consider sets $X = \{x_1, x_2\}$, $Y = \{y_1, y_2, y_3\}$, $Z = \{z_1, z_2, z_3\}$ with distribution $\omega \in \mathcal{D}(X \times Y)$ given by:

$$\tfrac{1}{4}|x_1, y_1\rangle + \tfrac{1}{4}|x_1, y_2\rangle + \tfrac{1}{6}|x_2, y_1\rangle + \tfrac{1}{6}|x_2, y_2\rangle + \tfrac{1}{6}|x_2, y_3\rangle$$

and $\rho \in \mathcal{D}(X \times Z)$ by:

$$\tfrac{1}{12}|x_1, z_1\rangle + \tfrac{1}{3}|x_1, z_2\rangle + \tfrac{1}{12}|x_1, z_3\rangle + \tfrac{1}{8}|x_2, z_1\rangle + \tfrac{1}{8}|x_2, z_2\rangle + \tfrac{1}{4}|x_2, z_3\rangle.$$

1 Check that ω and ρ have equal X-marginals.

2 Explicitly describe the extracted channels $c := \omega[0, 1 \,|\, 1, 0] \colon X \nrightarrow Y$ and $d := \omega[0, 1 \,|\, 1, 0] \colon X \nrightarrow Z$.

3 Show that the natural join $\omega \bowtie \rho \in \mathcal{D}(X \times Y \times Z)$ according to (7.22) is:

$$\begin{aligned}
&\tfrac{1}{24}|x_1, y_1, z_1\rangle + \tfrac{1}{6}|x_1, y_1, z_2\rangle + \tfrac{1}{24}|x_1, y_1, z_3\rangle \\
&\quad + \tfrac{1}{24}|x_1, y_2, z_1\rangle + \tfrac{1}{6}|x_1, y_2, z_2\rangle + \tfrac{1}{24}|x_1, y_2, z_3\rangle \\
&\quad + \tfrac{1}{24}|x_2, y_1, z_1\rangle + \tfrac{1}{24}|x_2, y_1, z_2\rangle + \tfrac{1}{12}|x_2, y_1, z_3\rangle \\
&\quad + \tfrac{1}{24}|x_2, y_2, z_1\rangle + \tfrac{1}{24}|x_2, y_2, z_2\rangle + \tfrac{1}{12}|x_2, y_2, z_3\rangle \\
&\quad + \tfrac{1}{24}|x_2, y_3, z_1\rangle + \tfrac{1}{24}|x_2, y_3, z_2\rangle + \tfrac{1}{12}|x_2, y_3, z_3\rangle.
\end{aligned}$$

7.6.7 Combining the two items of Theorem 7.6.3 gives a bijective correspondence between:

$$\frac{\Omega \in \mathcal{D}(\mathcal{D}(X)) \text{ with } \big|\mathit{supp}(\Omega)\big| = N}{\omega \in \mathcal{D}(X) \text{ and } c \colon X \nrightarrow N \text{ such that } c \gg\!= \omega \text{ has full support}}$$

Define this correspondence in detail and check that it is bijective.

7.6.8 Prove the following possibilitistic analogues of Theorem 7.6.3.

1 There is a bijective correspondence between:

$$\frac{R \in \mathcal{P}(X \times Y)}{U \in \mathcal{P}(X) \text{ with } f \colon U \to \mathcal{P}_*(Y)}$$

where $\mathcal{P}_*$ is used for the the subset of *non-empty* subsets.

2 For each number $N \geq 1$ there is a bijective correspondence:

$$\frac{A \in \mathcal{P}(\mathcal{P}(X)) \text{ with } |A| = N}{R \in \mathcal{P}(X \times N) \text{ with } \forall i \neq j.\, \exists x.\, \neg(R(x,i) \Leftrightarrow R(x,j))}$$

7.6.9 Prove that the equation

$$(\langle id, c\rangle \gg\!= \sigma)\big|_{p \otimes \mathbf{1}}[0,1 \,\big|\, 1,0] \;=\; c.$$

can be obtained both from Theorem 6.3.11 and from Theorem 7.6.4.

7.6.10 Prove that for a state $\omega \in \mathcal{D}(X \times Y)$ with full support one has:

$$\omega[1,0 \,\big|\, 0,1] \;=\; \Big(\omega[0,1 \,\big|\, 1,0]\Big)^{\dagger} \colon Y \nrightarrow X.$$

7.7 Disintegration and inversion in machine learning

This section illustrates the role of disintegration and Bayesian inversion in two fundamental techniques in machine learning, namely in naive Bayesian classification and in decision tree learning. These applications will be explained via examples from the literature.

7.7.1 Naive Bayesian classification

We illustrate the use of both disintegration and Bayesian inversion in an example of 'naive' Bayesian classification from [165]; we follow the analysis of [21]. Instead of trying to explain what a naive Bayesian classification is or does, we demonstrate via this example how it works. In the end, in Remark 7.7.1 we give a more general description.

Consider the table in Figure 7.1. It collects data about certain weather conditions and whether or not there is playing (outside). The question asked in [165] is: given this table, what can be said about the probability of playing if the outlook is *sunny*, the temperature is *cold*, the humidity is *high* and it is *windy*? This is a typical Bayesian update question, starting from (point) evidence. We will first analyse the situation in terms of channels.

We start by extracting the underlying spaces for the columns/categories in

Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Figure 7.1 Weather and play data, copied from [165].

the table in Figure 7.1. We choose obvious abbreviations for the entries in the table:

$$O = \{s, o, r\} \qquad T = \{h, m, c\} \qquad H = \{h, n\}. \qquad W = \{t, f\} \qquad P = \{y, n\}$$

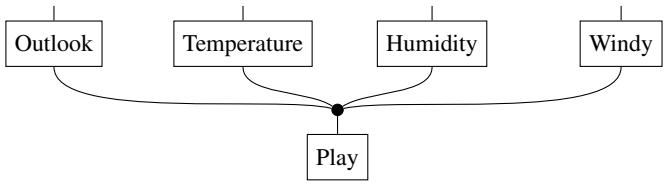
These sets are joined into a single product space:

$$S \coloneqq O \times T \times H \times W \times P.$$

It combines the five columns in Figure 7.1. The table itself can now be considered as a multiset in $\mathcal{M}(S)$ with 14 elements, each with multiplicity one. We will turn it immediately into an empirical distribution — formally via frequentist learning. It yields a distribution $\tau \in \mathcal{D}(S)$, with 14 entries, each with the same probability, written as:

$$\tau = \tfrac{1}{14}|s, h, h, f, n\rangle + \tfrac{1}{14}|s, h, h, t, n\rangle + \cdots + \tfrac{1}{14}|r, m, h, t, n\rangle.$$

We use a 'naive' Bayesian model in this situation, which means that we assume that all weather features are independent. This assumption can be visualised via the following string diagram:



(7.29)

This model oversimplifies the situation, but still it often leads to good (enough) outcomes.

We take the above perspective on the distribution τ, that is, we 'factorise' τ according to this string diagram (7.29). Obviously, the play state $\pi \in \mathcal{D}(P)$ is obtained as the last marginal:

$$\pi \coloneqq \tau[0,0,0,0,1] = \tfrac{9}{14}|y\rangle + \tfrac{5}{14}|n\rangle.$$

Next, we extract four channels c_O, c_T, c_H, c_W via appropriate disintegrations, from the Play column to the Outlook / Temperature / Humidity / Windy columns.

$$c_O \coloneqq \tau[1,0,0,0,0 \,|\, 0,0,0,0,1] \qquad c_H \coloneqq \tau[0,0,1,0,0 \,|\, 0,0,0,0,1]$$
$$c_T \coloneqq \tau[0,1,0,0,0 \,|\, 0,0,0,0,1] \qquad c_W \coloneqq \tau[0,0,0,1,0 \,|\, 0,0,0,0,1].$$

For instance the 'outlook' channel $c_O \colon P \nrightarrow O$ looks as follows.

$$c_O(y) = \tfrac{2}{9}|s\rangle + \tfrac{4}{9}|o\rangle + \tfrac{3}{9}|r\rangle \qquad c_O(n) = \tfrac{3}{5}|s\rangle + \tfrac{2}{5}|r\rangle. \tag{7.30}$$

It is analysed in greater detail in Exercise 7.7.1 below.

Now we can form the tuple channel of these extracted channels, called c in:

$$P \xrightarrow{\;c \coloneqq \langle c_O, c_T, c_H, c_W\rangle\;} O \times T \times H \times W$$

Recall the question that we started from: what is the probability of playing if the outlook is *sunny*, the temperature is *Cold*, the humidity is *High* and it is Windy? These features can be translated into an element (s, c, h, w) of the codomain $O \times T \times H \times W$ of this tuple channel — and thus into a point predicate. Hence our answer can be obtained by Bayesian inversion of the tuple channel, as:

$$\begin{aligned} c_\pi^\dagger(s,c,h,t) \overset{(6.6)}{=} \pi\big|_{c \ll \mathbf{1}_{(s,c,h,t)}} &= \tfrac{125}{611}|y\rangle + \tfrac{486}{611}|n\rangle \\ &\approx 0.2046|y\rangle + 0.7954|n\rangle. \end{aligned}$$

This corresponds to the probability 20.5% calculated in [165] — without any disintegration or Bayesian inversion.

The classification that we have just performed works via what it called a naive Bayesian *classifier*. In our set-up this classifier is the dagger channel:

$$O \times T \times H \times W \xrightarrow{\;c_\pi^\dagger\;} P$$

It predicts playing for a 4-tuple in $O \times T \times H \times W$.

In the end one can reconstruct a joint state with space S via the extracted channel, as graph:

$$\langle c_O, c_T, c_H, c_W, id\rangle \gg\!= \pi.$$

This state differs considerably from the original table/state τ. It shows that the shape (7.29) does not really fit the data that we have in Figure 7.1. But recall

that this approach is called naive. We shall soon look closer into such matters of shape in Section 7.9.

Remark 7.7.1. Now that we have seen the above illustration we can give a more abstract recipe of how to obtain a Bayesian classifier. The starting point is a joint state $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n \times Y)$ where $X_1, \ldots, X_n$ are the sets describing the 'input features' and Y is the set of 'target features' that are used in classification: its elements represent the different classes. The recipe involves the following steps, in which disintegration and Bayesian inversion play a prominent role.

1 Compute the prior classification probability $\pi := \omega[0, \ldots, 0, 1] \in \mathcal{D}(Y)$.

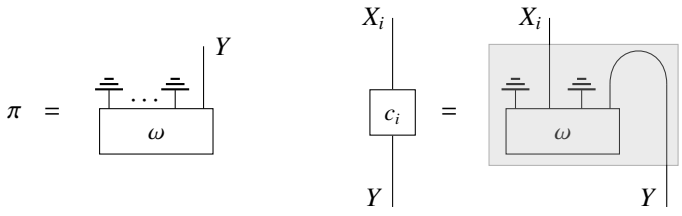
2 Extract n-channels $c_i \colon Y \nrightarrow X_i$ via disintegration:

$$\begin{aligned} c_1 &:= \omega[1, 0, \ldots, 0 \,\big|\, 0, \ldots, 0, 1] \colon Y \nrightarrow X_1 \\ c_2 &:= \omega[0, 1, 0 \ldots, 0 \,\big|\, 0, \ldots, 0, 1] \colon Y \nrightarrow X_2 \\ &\vdots \\ c_n &:= \omega[0, \ldots, 0, 1, 0 \,\big|\, 0, \ldots, 0, 1] \colon Y \nrightarrow X_n. \end{aligned}$$

3 Form the tuple channel $c := \langle c_1, \ldots, c_n \rangle \colon Y \nrightarrow X_1 \times \cdots \times X_n$.

4 Take the Bayesian inversion (dagger channel) $c_\pi^\dagger \colon X_1 \times \cdots \times X_n \nrightarrow Y$, as classifier. It gives for each n-tuple of input features $x_1, \ldots, x_n$ a distribution $c_\pi^\dagger(x_1, \ldots, x_n) \in \mathcal{D}(Y)$ on the set Y of classes (or target features). The distribution gives the probability of the tuple belonging to class $y \in Y$.

Graphically, these steps can be described as follows. First take:



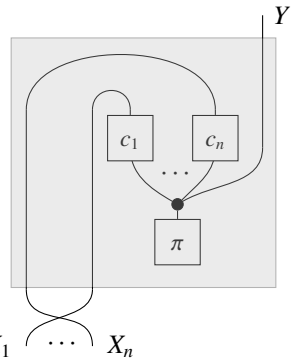
The classification channel $X_1 \times \cdots \times X_n \nrightarrow Y$ is then obtained via the dagger

Author	Thread	Length	Where read	User action
known	new	long	home	skip
unknown	new	short	work	read
unknown	follow-up	long	work	skip
known	follow-up	long	home	skip
known	new	short	home	read
known	follow-up	long	work	skip
unknown	follow-up	short	work	skip
unknown	new	short	work	read
known	follow-up	long	home	skip
known	new	long	work	skip
unknown	follow-up	short	home	skip
known	new	long	work	skip
known	follow-up	short	home	read
known	new	short	work	read
known	new	short	home	read
known	follow-up	short	work	read
known	new	short	home	read
unknown	new	short	work	read

Figure 7.2 Data about circumstances for reading or skipping of articles, copied from [140].

construction:

naive Bayesian classifier =



The twisting at the bottom happens implicitly when we consider the product $X_1 \times \cdots \times X_n$ as a single set and take the dagger wrt. this set.

7.7.2 Decision tree learning

We proceed as before, by first going trough an example from the literature, and then taking a step back to describe more abstractly what is going on. We start from a table of data in Figure 7.2, from [140, Fig. 7.1]. It describes user actions (read or skip) for articles that are "posted to a threaded discussion website depending on whether the author is known or not, whether the article started a

new thread or was a follow-up, the length of the article, and whether it is read at home or at work."

We formalise the table in Figure 7.2 via the following five sets, with elements corresponding in an obvious way to the entries in the table: k = known, u = unknown, *etc.*.

$$A = \{k, u\} \qquad T = \{n, f\} \qquad L = \{l, s\} \qquad W = \{h, w\} \qquad U = \{s, r\}.$$

We interprete the table as a joint distribution $\omega \in \mathcal{D}(A \times T \times L \times W \times U)$ with the same probability for each of the 18 entries in the table:

$$\omega = \tfrac{1}{18}\big|k, n, l, h, s\big\rangle + \tfrac{1}{18}\big|u, n, s, w, r\big\rangle + \cdots + \tfrac{1}{18}\big|u, n, s, w, r\big\rangle. \tag{7.31}$$

So far there is no real difference with the naive Bayesian classification example in the previous subsection. But the aim now is not *classification* but *deciding*: given a 4-tuple of input features $(x_1, x_2, x_3, x_4) \in A \times T \times L \times W$ we wish to decide as quickly as possible whether this tuple leads to a read or a skip action.

The way to take such a decision is to use a *decision tree* as described in Figure 7.3. It puts Length as dominant feature on top and tells that a long article is skipped immediately. Indeed, this is what we see in the table in Figure 7.2: we can quickly take this decision without inspecting the other features. If the article is short, the next most relevant feature, after Length, is Thread. Again we can see in Figure 7.2 that a short and new article is read. Finally, we see in Figure 7.3 that if the article is short and a follow-up, then it is read if the author is known, and skipped if the author is unkown. Apparently the location of the user (the Where feature) is irrelevant for the read/skip decision.

We see that such decision trees provide a (visually) clear and efficient method for reaching a decision about the target feature, starting from input features. The question that we wish to address here is: how to learn (derive) such a decision tree (in Figure 7.3) from the table (in Figure 7.2)?

The recipe (algorithm) for learning the tree involves iteratively going through the following three steps, acting on a joint state σ.

1 Check if we are done, that is, if the U-marginal (for User action) of the current joint state σ is a point state; if so, we are done and write this point state $1|x\rangle$ in a box as leaf in the tree.
2 If we are not done, determine the 'dominant' feature X for σ, and write X as a circle in the tree.
3 For each element of $x \in X$, update (and marginalise) σ with the point predicate $\mathbf{1}_x$ and re-start with step 1.

We shall describe these steps in more detail, starting from the state ω in (7.31), corresponding to the table in Figure 7.2.

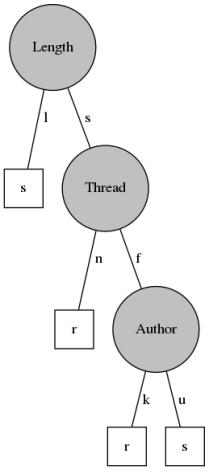


Figure 7.3 Decision tree for reading (r) or skipping (s) derived from the data in Figure 7.2.

1.1 The U-marginal $\omega_U \coloneqq \omega[0,0,0,0,1] \in \mathcal{D}(U)$ equals $\frac{1}{2}|r\rangle + \frac{1}{2}|s\rangle$, because there are equal numbers of read and skip actions in Figure 7.2. Since this is not a point state, we are not done.

1.2 We need to determine the dominant feature in ω. This is where disintegration comes in. We first extract four channels and states:

$$\begin{aligned}
c_A &\coloneqq \omega[0,0,0,0,1 \,|\, 1,0,0,0,0] : A \nrightarrow U \\
\omega_A &\coloneqq \omega[1,0,0,0,0] \in \mathcal{D}(A) \\
c_T &\coloneqq \omega[0,0,0,0,1 \,|\, 0,1,0,0,0] : T \nrightarrow U \\
\omega_T &\coloneqq \omega[0,1,0,0,0] \in \mathcal{D}(T) \\
c_L &\coloneqq \omega[0,0,0,0,1 \,|\, 0,0,1,0,0] : L \nrightarrow U \\
\omega_L &\coloneqq \omega[0,0,1,0,0] \in \mathcal{D}(L) \\
c_W &\coloneqq \omega[0,0,0,0,1 \,|\, 0,0,0,1,0] : W \nrightarrow U \\
\omega_W &\coloneqq \omega[0,0,0,1,0] \in \mathcal{D}(W).
\end{aligned}$$

Note that these channels go in the opposite direction with respect to the

approach for naive Bayesian classification. It is not hard to see that:

$$\begin{array}{lll} c_A(k) = \frac{1}{2}|s\rangle + \frac{1}{2}|r\rangle & c_A(u) = \frac{1}{2}|s\rangle + \frac{1}{2}|r\rangle & \omega_A = \frac{2}{3}|k\rangle + \frac{1}{3}|u\rangle \\ c_T(n) = \frac{3}{10}|s\rangle + \frac{7}{10}|r\rangle & c_T(f) = \frac{3}{4}|s\rangle + \frac{1}{4}|r\rangle & \omega_T = \frac{5}{9}|n\rangle + \frac{4}{9}|f\rangle \\ c_L(l) = 1|s\rangle & c_L(s) = \frac{2}{11}|s\rangle + \frac{9}{11}|r\rangle & \omega_L = \frac{7}{18}|l\rangle + \frac{11}{18}|s\rangle \\ c_W(h) = \frac{1}{2}|s\rangle + \frac{1}{2}|r\rangle & c_W(w) = \frac{1}{2}|s\rangle + \frac{1}{2}|r\rangle & \omega_W = \frac{4}{9}|h\rangle + \frac{5}{9}|w\rangle. \end{array}$$

In order to determine which of the input features A, T, L, W is most dominant we compute the expected entropy — sometimes called intrinsice value — for each of these features. Recall the Shannon entropy function $H\colon \mathcal{D}(X) \to \mathbb{R}_{\geq 0}$ from Exercise 4.1.9. Post-composing it with the above channels gives a factor, like $H \circ c_A\colon A \to \mathbb{R}_{\geq 0}$. It computes the entropy $H(c_A(x))$ for each $x \in A$. Hence we can compute its validity in the marginal state $\omega_A \in \mathcal{D}(A)$, giving the expected entropy. Explicitly:

$$\begin{aligned} \omega_A \models H \circ c_A &= \sum_{x \in A} \omega_A(x) \cdot H(c_A(x)) \\ &= \omega_A(k) \cdot H(c_A(k)) + \omega_A(u) \cdot H(c_A(u)) \\ &= \tfrac{2}{3} \cdot (\tfrac{1}{2} \cdot -\log(\tfrac{1}{2}) + \tfrac{1}{2} \cdot -\log(\tfrac{1}{2})) \\ &\qquad + \tfrac{1}{3} \cdot (\tfrac{1}{2} \cdot -\log(\tfrac{1}{2}) + \tfrac{1}{2} \cdot -\log(\tfrac{1}{2})) \\ &= \tfrac{2}{3} + \tfrac{1}{3} \\ &= 1. \end{aligned}$$

In the same way one computes the other expected entropies as:

$$\omega_T \models H \circ c_T = 0.85 \qquad \omega_L \models H \circ c_L = 0.42 \qquad \omega_W \models H \circ c_W = 1.$$

One then picks the lowest entropy value, which is 0.42, for feature / component L. Hence L = Length is the dominant feature at this first stage. Therefore it is put on top in the decision tree in Figure 7.3.

1.3 The set L has two elements, l for long and s for short. We update the current state ω with each of these, via suitably weakened point predicates $\mathbf{1}_l$ and $\mathbf{1}_s$, and marginalise out the L component. This gives new states for which we use the following *ad hoc* notation.

$$\begin{aligned} \omega/l &:= \omega\big|_{\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_l\otimes\mathbf{1}\otimes\mathbf{1}}[1,1,0,1,1] \in \mathcal{D}(A \times T \times W \times U) \\ \omega/s &:= \omega\big|_{\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_s\otimes\mathbf{1}\otimes\mathbf{1}}[1,1,0,1,1] \in \mathcal{D}(A \times T \times W \times U). \end{aligned}$$

We now go into a recursive loop and repeat the previous steps for both these states ω/l and ω/s. Notice that they are 'shorter' than ω, since they only have 4 components instead of 5, since we marginalised out the dominant component L.

2.1.1 In the l-branch we are now done, since the U-marginal of the l-update is a point state:

$$\omega/l\,[0,0,0,1] \;=\; 1|s\rangle.$$

It means that a long article is skipped immedately. This is indicated via the the l-box as (left) child of the Length node in the decision tree in Figure 7.3.

2.1.2 We continue with the s-branch. The U-marginal of ω/s is not a point state.

2.2.2 We will now have to determine the dominant feature in ω/s. We compute the three expected entropies for A, T, W as:

$$\begin{aligned}
\omega/s\,[1,0,0,0] &\models H \circ \omega/s\,[0,0,0,1 \,\big|\, 1,0,0,0] \;=\; 0.44\\
\omega/s\,[0,1,0,0] &\models H \circ \omega/s\,[0,0,0,1 \,\big|\, 0,1,0,0] \;=\; 0.36\\
\omega/s\,[0,0,1,0] &\models H \circ \omega/s\,[0,0,0,1 \,\big|\, 0,0,1,0] \;=\; 0.68.
\end{aligned}$$

The second value is the lowest, so that T = Thread is now the dominant feature. It is added as codomain node of the s-edge out Length in the decision tree in Figure 7.3.

2.3.2 The set T has two elements, n for new and f for follow-up. We take the corresponding updates:

$$\begin{aligned}
\omega/s/n &\coloneqq \omega/s\big|_{\mathbf{1}\otimes\mathbf{1}_n\otimes\mathbf{1}\otimes\mathbf{1}}[1,0,1,1] \;\in\; \mathcal{D}(A\times W\times U)\\
\omega/s/f &\coloneqq \omega/s\big|_{\mathbf{1}\otimes\mathbf{1}_f\otimes\mathbf{1}\otimes\mathbf{1}}[1,0,1,1] \;\in\; \mathcal{D}(A\times W\times U).
\end{aligned}$$

Then we enter new recursions with both these states.

3.1.1 In the n-branch we are done, since we find a point state as U-marginal:

$$\omega/s/n\,[0,0,1] \;=\; 1|r\rangle.$$

This settles the left branch under the Thread node in Figure 7.3.

3.1.2 The f-branch is not done, since the U-marginal of the state $\omega/s/f$ is not a point state.

3.2.2 We thus start computing expected entropies again, in order to find out which of the remaining input features A, T is dominant.

$$\begin{aligned}
\omega/s/f\,[1,0,0] &\models H \circ \omega/s/f\,[0,0,1 \,\big|\, 1,0,0] \;=\; 0\\
\omega/s/f\,[0,1,0] &\models H \circ \omega/s/f\,[0,0,1 \,\big|\, 0,1,0] \;=\; 1.
\end{aligned}$$

Hence the A feature is dominant, so that the Author node is added to the f-edge out of Thread in Figure 7.3.

3.3.2 The set A has two elements, k for known and and u for unknown. We form the corresponding two updates of the current state $\omega/s/f$.

$$\omega/s/f/k \coloneqq \omega/s/f\big|_{\mathbf{1}_k\otimes\mathbf{1}\otimes\mathbf{1}}[0,1,1] \;\in\; \mathcal{D}(W\times U)$$
$$\omega/s/f/u \coloneqq \omega/s/f\big|_{\mathbf{1}_u\otimes\mathbf{1}\otimes\mathbf{1}}[0,1,1] \;\in\; \mathcal{D}(W\times U).$$

The next cycle continues with these two states.

4.1.1 In the k-branch we are done since:

$$\omega/s/f/k\,[0,1] \;=\; 1|r\rangle.$$

4.1.2 Also in the u-branch we are done since:

$$\omega/s/f/u\,[0,1] \;=\; 1|s\rangle.$$

This gives the last two boxes, so that the decision tree in Figure 7.3 is finished.

There are many variations on the above learning algorithm for decision trees. The one that we just described is sometime called the 'classification' version, as in [91], since it works with discrete distributions. There is also a 'regression' version, for continuous distributions. The key part of the the above algorithm is deciding which feature is dominant (in step 2). We have described the so-called ID3 version from [142], which uses expected entropies (intrinsic values). Sometimes it is described in terms of 'gains', see [125], where in the above step 1.2 we can define for feature $X \in \{A, T, L, W\}$,

$$\mathit{gain}(X) \;\coloneqq\; H(\omega_U) - (\omega_X \models H \circ c_X).$$

One then looks for the feature with the highest gain. But one may as well look for the lowest expected entropy — given by the valdity expression after the minus sign — as we do above. There are alternatives to using gain, such as what is called 'gini', but that is out of scope.

To conclude, running the decision tree learning algorithm on the distribution associated with the weather and play table from Figure 7.1 — with Play as target feature — yields the decision tree in Figure 7.4, see also [165, Fig. 4.4].

Exercises

7.7.1 Write $\sigma \in \mathcal{M}(O \times T \times H \times W \times P)$ for the weather-play table in Figure 7.1, as multidimensional *multiset*.

1 Compute the marginal multiset $\sigma[1,0,0,0,1] \in \mathcal{M}(O \times P)$.

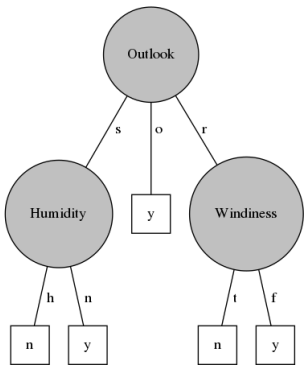


Figure 7.4 Decision tree for playing (y) or not (n) derived from the data in Figure 7.1.

2 Reorganise this marginalised multiset as a 2-dimensional table with only Outlook (horizontal) and Play (vertical) data, as given below, and check how this 'marginalised' table relates to the original one in Figure 7.1.

	Sunny	**Overcast**	**Rainy**
yes	2	4	3
no	3	0	2

3 Deduce a channel $P \nrightarrow O$ from this table, and compare it to the description (7.30) of the channel c_O given in Subsection 7.7.1 — see also Lemma 2.3.2 (and Proposition **??**).

4 Do the same for the marginal tables $\sigma[0,1,0,0,1]$, $\sigma[0,0,1,0,1]$, $\sigma[0,0,0,1,1]$, and the corresponding channels c_T, c_H, c_W in Subsection 7.7.1.

7.7.2 Continuing with the weather-play table σ, notice that instead of taking the dagger of the tuple of channels $\langle c_O, c_T, c_H, c_W \rangle \colon P \to O \times T \times H \times W$ in Example 7.7.1 we could have used instead the 'direct' disintegration:

$$O \times T \times H \times W \xrightarrow{\sigma[0,0,0,0,1 \,|\, 1,1,1,1,0]} P$$

Check that this gives a division-by-zero error.

7.7.3 Naive Bayesian classification, as illustrated in Subsection 7.7.1, is often used for classifying email messages as either 'spam' or 'ham' (not spam). One then looks for words which are typical for spam or ham. This exercise elaborates a standard small example.

Consider the following table of six words, together with the likelihoods of them being spam or ham.

	spam	**ham**
review	$1/4$	1
send	$3/4$	$1/2$
us	$3/4$	$1/2$
your	$3/4$	$1/2$
password	$1/2$	$1/2$
account	$1/4$	0

Lets write $S = \{s, h\}$ for the probability space for spam and ham, and, as usual $2 = \{0, 1\}$.

1 Translate the above table into six channels:

$$\text{review}, \text{send}, \text{us}, \text{your}, \text{password}, \text{account} \;:\; S \nrightarrow 2.$$

Write $c \colon S \nrightarrow 2^6 = 2 \times 2 \times 2 \times 2 \times 2 \times 2$ for the tuple channel:

$$c \coloneqq \langle \text{review}, \text{send}, \text{us}, \text{your}, \text{password}, \text{account} \rangle.$$

2 We wish to classify the message "review us now". Explain how it gets translated to the point predicate $\mathbf{1}_{(1,0,1,0,0,0)}$ on 2^6.

3 Show that:

$$\begin{aligned} c \mathrel{=\!\!\ll} \mathbf{1}_{(1,0,1,0,0,0)} &= (\text{review} \mathrel{=\!\!\ll} \mathbf{1}_1) \;\&\; (\text{send} \mathrel{=\!\!\ll} \mathbf{1}_0) \\ &\quad \;\&\; (\text{us} \mathrel{=\!\!\ll} \mathbf{1}_1) \;\&\; (\text{your} \mathrel{=\!\!\ll} \mathbf{1}_0) \\ &\quad \;\&\; (\text{password} \mathrel{=\!\!\ll} \mathbf{1}_0) \;\&\; (\text{account} \mathrel{=\!\!\ll} \mathbf{1}_0) \\ &= \tfrac{9}{2048} \cdot \mathbf{1}_s + \tfrac{1}{16} \cdot \mathbf{1}_h \end{aligned}$$

4 Let's assume a prior spam distribution $\omega = \frac{2}{3}|s\rangle + \frac{1}{3}|h\rangle$. Show that the posterior spam distribution for our message is:

$$\begin{aligned} c^\dagger_\omega(1,0,1,0,0,0) \;=\; \omega|_{c \mathrel{=\!\!\ll} \mathbf{1}_{(1,0,1,0,0,0)}} &= \tfrac{9}{73}|s\rangle + \tfrac{64}{73}|h\rangle \\ &\approx 0.1233|s\rangle + 0.8767|h\rangle. \end{aligned}$$

7.7.4 Show in detail that we get point states in items 2.1.1, 3.1.1, 4.1.1 and 4.1.2 in Subsection 7.7.2.

7.7.5 In Subsection 7.7.1 we classified the play distribution as $\frac{125}{611}|y\rangle + \frac{486}{611}|n\rangle$, given the input features (s, c, h, t). Check what the play decision is for these inputs in the decision tree in Figure 7.4.

7.7.6 As mentioned, the table in Figure 7.2 is copied from [140]. There the following two questions are asked: what can be said about reading/skipping for the two queries:

$$Q_1 = \text{(unknown, new, long, work)}$$
$$Q_2 = \text{(unknown, follow-up, short, home).}$$

1 Answer Q_1 and Q_2 via the decision tree in Figure 7.3.
2 Compute the distributions on U for both queries Q_1 and Q_2 via naive Bayesian classification.

(The outcome for Q_2 obtained via Bayesian classification is $\frac{3}{5}|r\rangle + \frac{2}{5}|s\rangle$; it gives reading the highest probability. This does not coincide with the outcome via the decision tree. Thus, one should be careful to rely on such classification methods for important decisions.)

7.8 Categorical aspects of Bayesian inversion

As suggested in the beginning of this section the 'dagger' of a channel — *i.e.* its Bayesian inversion — can also be described categorically. It turns out to be a special 'dagger' functor. Such reversal is quite common for non-deterministic computation, see Example 7.8.1 below. The fact that this same abstract structure exists for probabilistic computation demonstrates once again that Bayesian inversion is a canonical operation — and that category theory provides a useful language for making such similarities explicit.

This section goes a bit deeper into the category-theoretic aspects of probabilistic computation in general and of Bayesian inversion in particular. It is not essential for the rest of this book, but provides deeper insight into the underlying structures.

Example 7.8.1. Recall the category $\mathbf{Chan}(\mathcal{P})$ of non-deterministic computations. Its objects are sets X and its morphisms $f\colon X \nrightarrow Y$ are functions $f\colon X \to \mathcal{P}(X)$. The identity morphism $\mathit{unit}_X\colon X \nrightarrow X$ in $\mathbf{Chan}(\mathcal{P})$ is the singleton function $\mathit{unit}(x) = \{x\}$. Composition of $f\colon X \nrightarrow Y$ and $g\colon Y \nrightarrow Z$ is the function $g \odot f\colon X \nrightarrow Z$ given by:

$$(g \odot f)(x) = \{z \in Z \mid \exists y \in Y.\, y \in f(x) \text{ and } z \in g(y)\}.$$

It is not hard to see that $\odot$ is associative and has unit as identity element. In fact, this has already been proven more generally, in Lemma 1.8.3.

There are two aspects of the category $\mathbf{Chan}(\mathcal{P})$ that we wish to illustrate, namely (1) that it has an 'inversion' operation, in the form of a dagger functor,

and (2) that it is isomorphic to the category **Rel** of sets with relations between them (as morphisms). Probabilistic analogues of these two points will be described later.

1 We start from a very basic observation, namely that morphisms in $\mathbf{Chan}(\mathcal{P})$ can be reversed. There is a bijective correspondence, indicated by the double lines, between morphisms in $\mathbf{Chan}(\mathcal{P})$:

$$\frac{X \longrightarrow\!\!\!\!\circ\; Y}{Y \longrightarrow\!\!\!\!\circ\; X} \qquad \text{that is, between functions:} \qquad \frac{X \xrightarrow{f} \mathcal{P}(Y)}{Y \xrightarrow[g]{} \mathcal{P}(X)}$$

This correspondence sends $f\colon X \to \mathcal{P}(Y)$ to the function $f^\dagger\colon Y \to \mathcal{P}(X)$ with $f^\dagger(y) \coloneqq \{x \mid y \in f(x)\}$. Hence $y \in f(x)$ iff $x \in f^\dagger(y)$. Similarly one sends $g\colon Y \to \mathcal{P}(X)$ to $g^\dagger\colon X \to \mathcal{P}(Y)$ via $g^\dagger(x) \coloneqq \{y \mid x \in g(y)\}$. Clearly, $f^{\dagger\dagger} = f$ and $g^{\dagger\dagger} = g$.

It turns out that this dagger operation $(-)^\dagger$ interacts nicely with composition: one has $\mathit{unit}^\dagger = \mathit{unit}$ and also $(g \circ f)^\dagger = f^\dagger \circ g^\dagger$. This means that the dagger is functorial. It can be described as a functor $(-)^\dagger\colon \mathbf{Chan}(\mathcal{P}) \to \mathbf{Chan}(\mathcal{P})^{\mathrm{op}}$, which is the identity on objects: $X^\dagger = X$. The opposite $(-)^{\mathrm{op}}$ category is needed for this functor since it reverses arrows.

2 We write **Rel** for the category with sets X as objects and relations $R \subseteq X \times Y$ as morphisms $X \to Y$. The identity $X \to X$ is given by the equality relation $Eq_X \subseteq X \times X$, with $Eq_X = \{(x, x) \mid x \in X\}$. Composition of $R \subseteq X \times Y$ and $S \subseteq X \times Z$ is the 'relational' composition $S \bullet R \subseteq X \times Z$ given by:

$$S \bullet R \coloneqq \{(x, z) \mid \exists y \in Y.\, R(x, y) \text{ and } S(y, z)\}.$$

It is not hard to see that we get a category in this way.

There is a 'graph' functor $G\colon \mathbf{Chan}(\mathcal{P}) \to \mathbf{Rel}$, which is the identity on objects: $G(X) = X$. On a morphism $f\colon X \nrightarrow Y$, that is, on a function $f\colon X \to \mathcal{P}(Y)$, we define $G(f) \subseteq X \times Y$ to be $G(f) = \{(x, y) \mid x \in X, y \in f(x)\}$. Then: $G(\mathit{unit}_X) = Eq_X$ and $G(g \circ f) = G(f) \bullet G(f)$.

In the other direction there is also a functor $F\colon \mathbf{Rel} \to \mathbf{Chan}(\mathcal{P})$, which is again the identity on objects: $F(X) = X$. On a morphism $R\colon X \to Y$ in **Rel**, that is, on a relation $R \subseteq X \times Y$, we define $F(R)\colon X \to \mathcal{P}(Y)$ as $F(R)(x) = \{y \mid R(x, y)\}$. This F preserves identities and composition.

These two functors G and F are each other's inverses, in the sense that:

$$F \circ G = \mathit{id}\colon \mathbf{Chan}(\mathcal{P}) \to \mathbf{Chan}(\mathcal{P}) \quad \text{and} \quad G \circ F = \mathit{id}\colon \mathbf{Rel} \to \mathbf{Rel}.$$

This establishes an isomorphism $\mathbf{Chan}(\mathcal{P}) \cong \mathbf{Rel}$ of categories.

Interestingly, **Rel** is also a dagger category, via the familar operation of

reversal of relations: for $R \subseteq X \times Y$ one can form $R^\dagger \subseteq Y \times X$ via $R^\dagger(y,x) = R(x,y)$. This yields a functor $(-)^\dagger \colon \mathbf{Rel} \to \mathbf{Rel}^{\mathrm{op}}$, obviously with $(-)^{\dagger\dagger} = id$.

Moreover, the above functors G and F commute with the daggers of $\mathbf{Chan}(\mathcal{P})$ and $\mathbf{Rel}$, in the sense that:

$$G(f^\dagger) = G(f)^\dagger \qquad \text{and} \qquad F(R^\dagger) = F(R)^\dagger.$$

We shall prove the first equation and leave the second one to the interested reader. The proof is obtained by carefully unpacking the right definition at each stage. For a function $f \colon X \to \mathcal{P}(Y)$ and elements $x \in X$, $y \in Y$,

$$G(f^\dagger)(y,x) \Leftrightarrow x \in f^\dagger(y) \Leftrightarrow y \in f(x) \Leftrightarrow G(f)(x,y) \Leftrightarrow G(f)^\dagger(y,x).$$

We now move from non-deterministic to probabilistic computation. Our aim is to obtain analogous results, namely inversion in the form of a dagger functor on a category of probabilistic channels, and an isomorphism of this category with a category of probabilistic relations. One may expect that these results hold for the category $\mathbf{Chan}(\mathcal{D})$ of probabilistic channels. But the situation is a bit more subtle. Recall from the previous section that the dagger (Bayesian inversion) $c^\dagger_\omega \colon Y \nrightarrow X$ of a probabilistic channel $c \colon X \to Y$ requires a state $\omega \in \mathcal{D}(X)$ on the domain. In order to conveniently deal with this situation we incorporate these states ω into the objects of our category. We follow [24] and denote this category as **Krn**; its morphisms are 'kernels'.

Definition 7.8.2. The category **Krn** of kernels has:

- objects: pairs (X, σ) where X is a finite set and $\sigma \in \mathcal{D}(X)$ is a distribution on X with full support;
- morphisms: $f \colon (X,\sigma) \to (Y,\tau)$ are probabilistic channels $f \colon X \nrightarrow Y$ with $f \gg\!= \sigma = \tau$.

Identity maps $(X,\sigma) \to (X,\sigma)$ in **Krn** are identity channels $unit \colon X \nrightarrow X$, given by $unit(x) = 1|x\rangle$, which we write simply as id. Composition in **Krn** is ordinary composition $\circ$ of channels.

Theorem 7.8.3. *Bayesian inversion forms a dagger functor* $(-)^\dagger \colon \mathbf{Krn} \to \mathbf{Krn}^{\mathrm{op}}$ *which is the identity on objects and which sends:*

$$\Big((X,\sigma) \xrightarrow{f} (Y,\tau)\Big) \longmapsto \Big((Y,\tau) \xrightarrow{f^\dagger_\sigma} (X,\sigma)\Big)$$

This functor is its own inverse: $f^{\dagger\dagger} = f$.

Proof. We first have to check that the dagger functor is well-defined, *i.e.* that

the above mapping yields another morphism in **Krn**. This follows from Proposition 6.6.4 (1):

$$f^\dagger_\sigma \gg\!= \tau \;=\; f^\dagger_\sigma \gg\!= (f \gg\!= \sigma) \;=\; \sigma.$$

Aside: this does *not* mean that $f^\dagger \circ f = id$.

We next show that the dagger preserves identities and composition. We use the concrete formulation of Bayesian inversion of Definition 6.6.1. A string-diagrammatic proof is possible, see below. The identity map $id \colon (X, \sigma) \to (X, \sigma)$ in **Krn** satisfies:

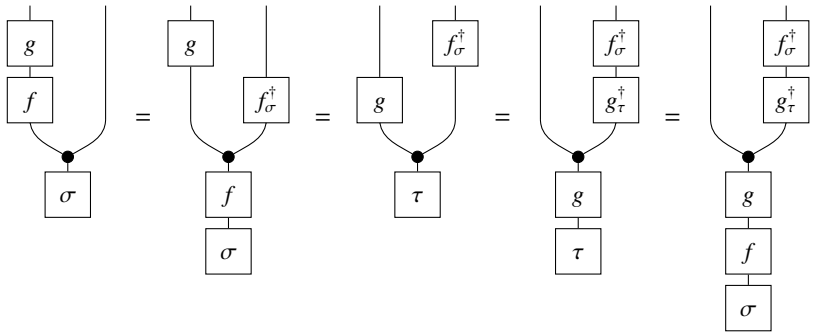
$$\begin{aligned} id^\dagger_\sigma(x) \;=\; \sigma|_{id \mathrel{=\!\!\ll} \mathbf{1}_x} \;&=\; \sigma|_{\mathbf{1}_x} \;=\; 1|x\rangle \qquad \text{see Exercise 6.1.1} \\ &= id(x). \end{aligned}$$

For morphisms $(X, \sigma) \xrightarrow{f} (Y, \tau) \xrightarrow{g} (Z, \rho)$ in **Krn** we have, for $z \in Z$,

$$\begin{aligned} (g \circ f)^\dagger_\sigma(z) \;=\; \sigma|_{(g \circ f) \mathrel{=\!\!\ll} \mathbf{1}_z} \;&= \sigma|_{f \mathrel{=\!\!\ll} (g \mathrel{=\!\!\ll} \mathbf{1}_z)} \\ &= f^\dagger_\sigma \gg\!= ((f \gg\!= \sigma)|_{g \mathrel{=\!\!\ll} \mathbf{1}_z}) \quad \text{by Theorem 6.6.3 (1)} \\ &= f^\dagger_\sigma \gg\!= (\tau|_{g \mathrel{=\!\!\ll} \mathbf{1}_z}) \\ &= f^\dagger_\sigma \gg\!= g^\dagger_\tau(z) \\ &= (f^\dagger_\sigma \circ g^\dagger_\tau)(z). \end{aligned}$$

Finally we have $f^\dagger_\sigma)^\dagger_\tau = f$ by Proposition 6.6.4 (2). ❑

One can also prove this result via equational reasoning with string diagrams. For instance preservation of composition $\circ$ by the dagger follows by uniqueness from:



We now turn to probabilistic relations, with the goal of finding a category of such relations that is isomorphic to **Krn**. For this purpose we use what are called *couplings*. The definition that we use below differs only in inessential ways from the formulation that we used in Subsection 4.5.1 for the Wasserstein distance. For more information couplings, see *e.g.* [11].

Definition 7.8.4. We introduce a category **Cpl** of couplings with the same objects as **Krn**. A morphism $(X, \sigma) \to (Y, \tau)$ in **Cpl** is a joint state $\varphi \in \mathcal{D}(X \times Y)$ with $\varphi[1, 0] = \sigma$ and $\varphi[0, 1] = \tau$. Such a distribution which marginalises to σ and τ is called a coupling between σ and τ.

Composition of $\varphi \colon (X, \sigma) \to (Y, \tau)$ and $\psi \colon (Y, \tau) \to (Z, \rho)$ is the distribution $\psi \bullet \varphi \in \mathcal{D}(X \times Z)$ defined as:

$$\begin{aligned} \psi \bullet \varphi &:= \langle \varphi[1, 0 \,|\, 0, 1], \psi[0, 1 \,|\, 1, 0] \rangle \gg\!= \tau \\ &= \sum_{x \in X,\, z \in Z} \left(\sum_{y \in Y} \frac{\varphi(x, y) \cdot \psi(y, z)}{\tau(y)} \right) |\, x, z \,\rangle. \end{aligned} \tag{7.32}$$

The identity coupling $Eq_{(X,\sigma)} \colon (X, \sigma) \to (X, \sigma)$ is the distribution $\Delta \gg\!= \sigma$.

The essence of the following result is due to [24], but there it occurs in slightly different form, namely in a setting of continuous probability. Here it is translated to the discrete situation.

Theorem 7.8.5. *Couplings as defined above indeed form a category* **Cpl**.

1 *This category carries a dagger functor* $(-)^\dagger \colon \mathbf{Cpl} \to \mathbf{Cpl}^{\mathrm{op}}$ *which is the identity on objects; on morphisms it is defined via swapping:*

$$\Big((X, \sigma) \xrightarrow{\varphi} (Y, \tau) \Big)^\dagger := \Big((Y, \tau) \xrightarrow{\langle \pi_2, \pi_1 \rangle \gg\!= \varphi} (X, \sigma) \Big).$$

More concretely, this dagger is defined by swapping arguments, as in: $\varphi^\dagger = \sum_{x,y} \varphi(x, y) |y, x\rangle$.

2 *There is an isomorphism of categories* $\mathbf{Krn} \cong \mathbf{Cpl}$*, in one direction by taking the graph of a channel, and in the other direction by disintegration. This isomorphism commutes with the daggers on the two categories.*

Proof. We first need to prove that $\psi \bullet \varphi$ is a distribution:

$$\begin{aligned} \sum_{x,z} (\psi \bullet \varphi)(x, z) &= \sum_{x,y,z} \frac{\varphi(x, y) \cdot \psi(y, z)}{\tau(y)} \\ &= \sum_{y,z} \frac{\big(\sum_x \varphi(x, y) \big) \cdot \psi(y, z)}{\tau(y)} \\ &= \sum_{y,z} \frac{\varphi[0, 1](y) \cdot \psi(y, z)}{\tau(y)} \\ &= \sum_{y,z} \frac{\tau(y) \cdot \psi(y, z)}{\tau(y)} = \sum_{y,z} \psi(y, z) = 1. \end{aligned}$$

We leave it to the reader to check that $Eq_{(X,\sigma)} = \Delta \gg\!= \sigma$ is neutral element for $\bullet$. We do verify that $\bullet$ is associative — and thus that **Cpl** is indeed a category.

Let $\varphi\colon (X,\sigma) \to (Y,\tau)$, $\psi\colon (Y,\tau) \to (Z,\rho)$, $\chi\colon (Z,\rho) \to (W,\kappa)$ be morphisms in **Cpl**. Then:

$$\begin{aligned}\Big(\chi \bullet (\psi \bullet \varphi)\Big)(x,w) &= \sum_z \frac{(\psi \bullet \varphi)(x,z) \cdot \chi(z,w)}{\rho(z)} \\ &= \sum_{y,z} \frac{\varphi(x,y) \cdot \psi(y,z) \cdot \chi(z,w)}{\tau(y) \cdot \rho(z)} \\ &= \sum_y \frac{\varphi(x,y) \cdot (\chi \bullet \psi)(y,w)}{\tau(y)} \;=\; \Big((\chi \bullet \psi) \bullet \varphi\Big)(x,w).\end{aligned}$$

We turn to the dagger. It is obvious that $(-)^{\dagger\dagger}$ is the identity functor, and also that $(-)^\dagger$ preserves identity maps. It also preserves composition in **Cpl** since:

$$\begin{aligned}(\psi \bullet \varphi)^\dagger(z,x) &= (\psi \bullet \varphi)(x,z) \\ &= \sum_y \frac{\varphi(x,y) \cdot \psi(y,z)}{\tau(y)} \\ &= \sum_y \frac{\psi^\dagger(z,y) \cdot \varphi^\dagger(y,x)}{\tau(y)} \;=\; (\varphi^\dagger \bullet \psi^\dagger)(z,x).\end{aligned}$$

The graph operation *gr* on channels from Definition 2.5.6 gives rise to an identity-on-objects 'graph' functor $G\colon \mathbf{Krn} \to \mathbf{Cpl}$ via:

$$G\Big((X,\sigma) \xrightarrow{f} (Y,\tau)\Big) \coloneqq \Big((Y,\tau) \xrightarrow{gr(\sigma,f)} (X,\sigma)\Big).$$

where $gr(\sigma,f) = \langle id, f\rangle \gg\!= \sigma$. This yields a functor since:

$$\begin{aligned}G(id_{(X,\sigma)}) &= \langle id, id\rangle \gg\!= \sigma \\ &= Eq_{(X,\sigma)} \\ G(g \circ f)(x,z) &= gr(\sigma, g \circ f)(x,z) \\ &= \sigma(x) \cdot (g \circ f)(x)(z) \\ &= \sum_y \sigma(x) \cdot f(x)(y) \cdot g(y)(z) \\ &= \sum_y \frac{\sigma(x) \cdot f(x)(y) \cdot \tau(y) \cdot g(y)(z)}{\tau(y)} \\ &= \sum_y \frac{gr(\sigma,f)(x,y) \cdot gr(\tau,g)(y,z)}{\tau(y)} \\ &= (gr(\tau,g) \bullet gr(\sigma,f))(x,z) \\ &= (G(g) \bullet G(f))(x,z).\end{aligned}$$

In the other direction we define a functor $F\colon \mathbf{Cpl} \to \mathbf{Krn}$ which is the identity on objects and use disintegration on morphisms: for $\varphi\colon (X,\sigma) \to (Y,\tau)$ in **Cpl** we get a channel $F(\varphi) \coloneqq dis_1(\varphi) = \varphi[0,1 \,\big|\, 1,0]\colon X \nrightarrow Y$ which satisfies, by construction (7.20):

$$\varphi \;=\; gr(\varphi[1,0], dis_1(\varphi)) \;=\; gr(\sigma, F(\varphi)) \;=\; GF(\varphi).$$

Moreover, $F(\varphi)$ is a morphism $(X, \sigma) \to (Y, \tau)$ in **Krn** since:

$$\begin{aligned} F(\varphi) \gg\!= \sigma &= dis_1(\varphi) \gg\!= (\varphi[1,0]) \\ &= \Big(gr(\varphi[1,0], dis_1(\varphi))\Big)[0,1] \overset{(7.19)}{=} \varphi[0,1] \;=\; \tau. \end{aligned}$$

We still need to prove that F preserves identities and composition. This follows by uniqueness of disintegration:

$$\begin{aligned} &\langle id, F(Eq_{(X,\sigma)})\rangle \gg\!= \sigma \\ &\quad= Eq_{(X,\sigma)} && \text{by definition} \\ &\quad= \Delta \gg\!= \sigma \\ &\quad= \langle id, id\rangle \gg\!= \sigma \\ &\langle id, F(\psi \bullet \varphi)\rangle \gg\!= \sigma \\ &\quad= \psi \bullet \varphi && \text{by definition} \\ &\overset{(7.32)}{=} \langle \varphi[1,0\,|\,0,1], \psi[0,1\,|\,1,0]\rangle \gg\!= \tau \\ &\quad= (id \otimes \psi[0,1\,|\,1,0]) \gg\!= \Big(\langle \varphi[1,0\,|\,0,1], id\rangle \gg\!= (\varphi[1,0\,|\,0,1] \gg\!= \sigma)\Big) \\ &\quad= (id \otimes F(\psi)) \gg\!= \Big(\langle F(\varphi)^\dagger, id\rangle \gg\!= (F(\varphi) \gg\!= \sigma)\Big) && \text{by Exercise 7.6.10} \\ &\overset{(7.26)}{=} (id \otimes F(\psi)) \gg\!= (\langle id, F(\varphi)\rangle \gg\!= \sigma) \\ &\quad= \langle id, F(\psi) \circ F(\varphi)\rangle \gg\!= \sigma. \end{aligned}$$

We have seen that, by construction, $G \circ F$ is the identity functor on the category **Cpl**. In the other direction we also have $F \circ G = id : \mathbf{Krn} \to \mathbf{Krn}$. This follows directly from uniqueness of disintegration.

In the end we like to show that the functors G and F commute with the daggers, on kernels and couplings. This works as follows. First, for $f\colon (X, \sigma) \to (Y, \tau)$ we have:

$$\begin{aligned} G(f^\dagger)(y,x) \;=\; gr(\tau, f^\dagger)(y,x) &= \tau(y) \cdot f^\dagger_\sigma(y)(x) \\ &= (f \gg\!= \sigma)(y) \cdot \frac{\sigma(x) \cdot f(x)(y)}{(f \gg\!= \sigma)(y)} \\ &= \sigma(x) \cdot f(x)(y) \\ &= gr(\sigma, f)(x,y) \\ &= G(f)(x,y) \;=\; G(f)^\dagger(y,x). \end{aligned}$$

Next, for $\varphi\colon (X,\sigma) \to (Y,\tau)$ in **Cpl**,

$$\begin{array}{rcll} F(\varphi^\dagger)(y)(x) & = & dis_1(\varphi^\dagger)(y)(x) & \\ & \overset{(7.21)}{=} & \dfrac{\varphi^\dagger(y,x)}{\varphi^\dagger[1,0](y)} & \\ & = & \dfrac{\varphi(x,y)}{\varphi[0,1](y)} & \\ & = & \varphi[1,0 \,\big|\, 0,1](y)(x) & \\ & = & (\varphi[0,1 \,\big|\, 1,0])^\dagger(y)(x) & \text{by Exercise 7.6.10} \\ & = & F(\varphi)^\dagger(y)(x). & \end{array}$$ ❑

We have done a lot of work in order to be able to say that **Krn** and **Cpl** are isomorphic dagger categories, or, more informally, that there is a one-one correspondence between probabilistic computations (channels) and probabilistic relations.

Exercises

7.8.1 Prove in the context of the powerset channels of Example 7.8.1 that:

1 $unit^\dagger = unit$ and $(g \circ f)^\dagger = f^\dagger \circ g^\dagger$.
2 $G(unit_X) = Eq_X$ and $G(g \circ f) = G(f) \bullet G(f)$.
3 $F(Eq_X) = unit_X$ and $F(S \bullet R) = F(S) \circ F(R)$.
4 $(S \bullet R)^\dagger = R^\dagger \bullet S^\dagger$.
5 $F(R^\dagger) = F(R)^\dagger$.

7.8.2 Give a string diagrammatic proof of the property $f^{\dagger\dagger} = f$ in Theorem 7.8.3. Prove also that $(f \otimes g)^\dagger = f^\dagger \otimes g^\dagger$.

7.8.3 Prove the equation = in the definition (7.32) of composition $\bullet$ in the category **Cpl**. Give also a string diagrammatic description of $\bullet$.

7.9 Factorisation of joint states

Earlier, in Subsection 2.5.1, we have called a binary joint state non-entwined when it is the product of its marginals. This can be seen as an intrinsic property of the state, which we will express in terms of a string diagram, called its *shape*. For instance, the state:

$$\omega \;=\; \tfrac{1}{4}|a,b\rangle + \tfrac{1}{2}|a,b^\perp\rangle + \tfrac{1}{12}|a^\perp,b\rangle + \tfrac{1}{6}|a^\perp,b^\perp\rangle$$

is non-entwined: it is the product of its marginals $\omega[1,0] = \frac{3}{4}|a\rangle + \frac{1}{4}|a^{\perp}\rangle$ and $\omega[0,1] = \frac{1}{3}|b\rangle + \frac{2}{3}|b^{\perp}\rangle$. We will formulate this as:

ω has *shape*
or as: and write this as: $\omega \models\!\approx$.
ω *factorises* as

We shall give a formal definition of $\models\!\approx$ below, but at this stage it suffices to read $\sigma \models\!\approx S$, for a state σ and a string diagram S, as: there is an interpretation of the boxes in S such that $\sigma = [\![S]\!]$.

In the above case of $\omega \models\!\approx$ we obtain $\omega = \omega_1 \otimes \omega_2$, for some state ω_1 that interpretes the box on the left, and some ω_2 interpreting the box on the right. But then:

$$\omega[1,0] = (\omega_1 \otimes \omega_2)[1,0] = \omega_1.$$

Similarly, $\omega[0,1] = \omega_2$. Thus, in this case the interpretations of the boxes are uniquely determined, namely as first and second marginal of ω.

We conclude that non-entwinedness of an arbitrary binary joint state ω can be expressed as: $\omega \models\!\approx$. Here we are interested in similar intrinsic 'shape' properties of states and channels that can be expressed via string diagrams. These matters are often discussed in the literature in terms of (conditional) independencies. Here we prefer to use shapes instead of independencies since they are more expressive.

In general there may be several interpretations of a string diagram (as shape). Consider for instance:

$$\tfrac{14}{25}|H\rangle + \tfrac{11}{25}|T\rangle \models\!\approx$$

This state has this form in multiple ways, for instance as:

$$c_1 \gg\!= \sigma_1 = \tfrac{14}{25}|H\rangle + \tfrac{11}{25}|T\rangle = c_2 \gg\!= \sigma_2$$

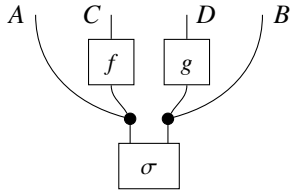
for:

$$\begin{array}{rcl@{\qquad}rcl} \sigma_1 &=& \frac{1}{5}|1\rangle + \frac{4}{5}|0\rangle & & \sigma_2 &=& \frac{2}{5}|1\rangle + \frac{3}{5}|0\rangle \\ c_1(1) &=& \frac{4}{5}|H\rangle + \frac{1}{5}|T\rangle & \text{and} & c_2(1) &=& \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle \\ c_1(0) &=& \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle & & c_2(0) &=& \frac{3}{5}|H\rangle + \frac{2}{5}|T\rangle. \end{array}$$

We note that is not an accessible string diagram: the wire inbetween the two boxes cannot be accessed from the outside. If these wires are accessible, then we can access the individual boxes of a string diagram and use disintegration to compute them. We illustrate how this works.

Example 7.9.1. Consider two-element sets $A = \{a, a^\perp\}$, $B = \{b, b^\perp\}$, $C = \{c, c^\perp\}$ and $D = \{d, d^\perp\}$ and an (accessible) string diagram S of the form:

S =



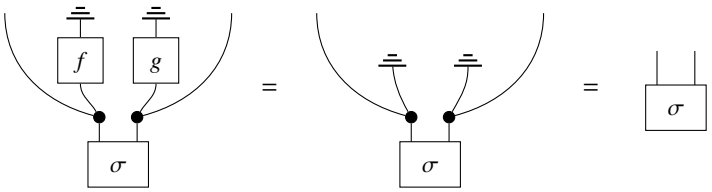
(7.33)

Now suppose we have a joint state $\omega \in \mathcal{D}(A \times C \times D \times B)$ given by:

$$\begin{aligned}\omega \;=\; & \tfrac{1}{25}|a,c,d,b\rangle + \tfrac{9}{50}|a,c,d,b^\perp\rangle + \tfrac{3}{50}|a,c,d^\perp,b\rangle + \tfrac{1}{50}|a,c,d^\perp,b^\perp\rangle \\ & + \tfrac{1}{25}|a,c^\perp,d,b\rangle + \tfrac{9}{50}|a,c^\perp,d,b^\perp\rangle + \tfrac{3}{50}|a,c^\perp,d^\perp,b\rangle \\ & + \tfrac{1}{50}|a,c^\perp,d^\perp,b^\perp\rangle + \tfrac{3}{125}|a^\perp,c,d,b\rangle + \tfrac{9}{500}|a^\perp,c,d,b^\perp\rangle \\ & + \tfrac{9}{250}|a^\perp,c,d^\perp,b\rangle + \tfrac{1}{500}|a^\perp,c,d^\perp,b^\perp\rangle + \tfrac{12}{125}|a^\perp,c^\perp,d,b\rangle \\ & + \tfrac{9}{125}|a^\perp,c^\perp,d,b^\perp\rangle + \tfrac{18}{125}|a^\perp,c^\perp,d^\perp,b\rangle + \tfrac{1}{125}|a^\perp,c^\perp,d^\perp,b^\perp\rangle\end{aligned}$$

We ask ourselves: does $\omega \models\!\approx S$ hold? More specifically, can we somehow obtain interpretations of the boxes σ, f and g in S so that $\omega = [\![S]\!]$? We shall show that by appropriately using marginalisation and disintegration we can 'factorise' this joint state according to the above string diagram.

First we can obtain the state $\sigma \in \mathcal{D}(A \times B)$ by discarding the C, D outputs in the middle, as in:

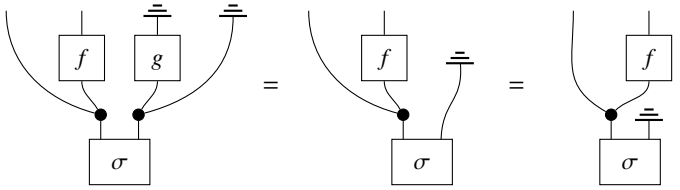


We can thus compute σ as:

$$\begin{aligned}\sigma \;&=\; \omega[1,0,0,1] \\ &=\; (\textstyle\sum_{u,v} \omega(a,u,v,b))|a,b\rangle + (\sum_{u,v} \omega(a,u,v,b^\perp))|a,b^\perp\rangle \\ &\qquad + (\textstyle\sum_{u,v} \omega(a^\perp,u,v,b))|a^\perp,b\rangle + (\sum_{u,v} \omega(a^\perp,u,v,b^\perp))|a^\perp,b^\perp\rangle \\ &=\; (\tfrac{1}{25} + \tfrac{3}{50} + \tfrac{1}{25} + \tfrac{3}{50})|a,b\rangle + (\tfrac{9}{50} + \tfrac{1}{50} + \tfrac{9}{50} + \tfrac{1}{50})|a,b^\perp\rangle \\ &\qquad + (\tfrac{3}{125} + \tfrac{9}{250} + \tfrac{12}{125} + \tfrac{18}{125})|a^\perp,b\rangle + (\tfrac{9}{500} + \tfrac{1}{500} + \tfrac{9}{125} + \tfrac{1}{125})|a^\perp,b^\perp\rangle \\ &=\; \tfrac{1}{5}|a,b\rangle + \tfrac{2}{5}|a,b^\perp\rangle + \tfrac{3}{10}|a^\perp,b\rangle + \tfrac{1}{10}|a^\perp,b^\perp\rangle.\end{aligned}$$

Next we concentrate on the channels f and g, from A to C and from B to D. We first illustrate how to restrict the string diagram to the relevant part via

marginalisation. For f we concentrate on:



The string diagram on the right tells us that we can obtain f via disintegration from the marginal $\omega[1,1,0,0]$, using that extracted channels are unique, in diagrams of this form, see (7.19). In the same way one obtains g from the marginal $\omega[0,0,1,1]$. Thus:

$$\begin{aligned}
f &= \omega[0,1,0,0 \,\big|\, 1,0,0,0] \\
&= \begin{cases} a \mapsto \dfrac{\sum_{v,y} \omega(a,c,v,y)}{\sum_{u,v,y} \omega(a,u,v,y)}|c\rangle + \dfrac{\sum_{v,y} \omega(a,c^\perp,v,y)}{\sum_{u,v,y} \omega(a,u,v,y)}|c^\perp\rangle \\ a^\perp \mapsto \dfrac{\sum_{v,y} \omega(a^\perp,c,v,y)}{\sum_{u,v,y} \omega(a^\perp,u,v,y)}|c\rangle + \dfrac{\sum_{v,y} \omega(a^\perp,c^\perp,v,y)}{\sum_{u,v,y} \omega(a^\perp,u,v,y)}|c^\perp\rangle \end{cases} \\
&= \begin{cases} a \mapsto \frac{1}{2}|c\rangle + \frac{1}{2}|c^\perp\rangle \\ a^\perp \mapsto \frac{1}{5}|c\rangle + \frac{4}{5}|c^\perp\rangle \end{cases} \\
g &= \omega[0,0,1,0 \,\big|\, 0,0,0,1] \\
&= \begin{cases} b \mapsto \dfrac{\sum_{x,u} \omega(x,u,d,b)}{\sum_{x,u,v} \omega(x,u,v,b)}|d\rangle + \dfrac{\sum_{x,u} \omega(x,u,d^\perp,b)}{\sum_{x,u,v} \omega(x,u,v,b)}|d^\perp\rangle \\ b^\perp \mapsto \dfrac{\sum_{x,u} \omega(x,u,d,b^\perp)}{\sum_{x,u,v} \omega(x,u,v,b^\perp)}|d\rangle + \dfrac{\sum_{x,v} \omega(x,u,d^\perp,b^\perp)}{\sum_{x,u,v} \omega(x,u,v,b^\perp)}|d^\perp\rangle \end{cases} \\
&= \begin{cases} b \mapsto \frac{2}{5}|d\rangle + \frac{3}{5}|d^\perp\rangle \\ b^\perp \mapsto \frac{9}{10}|d\rangle + \frac{1}{10}|d^\perp\rangle. \end{cases}
\end{aligned}$$

At this stage one can check that the joint state ω can be reconstructed from these extracted state and channels, namely as:

$$[\![S]\!] = (id \otimes f \otimes g \otimes id) \gg\!= ((\Delta \otimes \Delta) \gg\!= \sigma) = \omega.$$

This proves $\omega \models\!\approx S$.

We now come to the definition of $\models\!\approx$. We will use it for channels, and not just for states, as used above.

Definition 7.9.2. Let $A_1, \ldots, A_n, B_1, \ldots, B_m$ be finite sets. Consider a channel $c \colon A_1 \times \cdots \times A_n \nrightarrow B_1 \times \cdots \times B_m$ and a string diagram $S \in SD(\Sigma)$ over signature Σ, with domain $[A_1, \ldots, A_n]$ and codomain $[B_1, \ldots, B_m]$. We say that c and S have the same type.

In this situation we write:

$$c \models\!\approx S$$

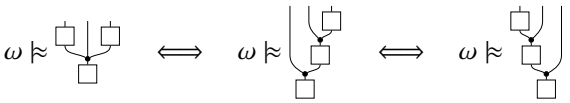
if there is an interpretation of the boxes in Σ such that $c = [\![S]\!]$.

We then say that c has *shape* S and also that c *factorises* according to S. Alternative terminology is: c is Markov relative to S, or: S represents c. Given c and S, finding an interpretation of Σ such that $c = [\![S]\!]$ is called a factorisation problem. It may have no, exactly one, or more than one solution (interpretation), as we have seen above.

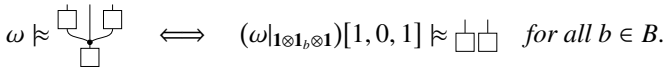
The following illustration is a classical one, showing how seemingly different shapes are related. It is often used to describe conditional independence — in this case of A, C, given B.

Theorem 7.9.3. *Let a state $\omega \in \mathcal{D}(A \times B \times C)$ have full support.*

1 There are equivalences:



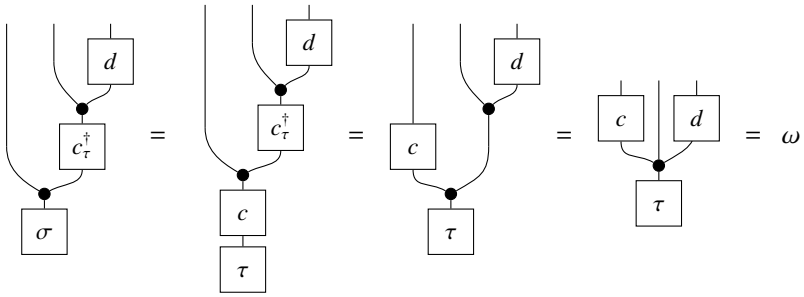
2 These equivalences can be extended as:



The string diagrams on the left in item (1) is often called a fork; the other two are called a chain. In item (2) this shape is related to non-entwinedness, pointwise.

Proof. 1 Since ω has full support, so have all its marginals, see Exercise 7.5.1. This allows us to perform all disintegrations below. We start on the left-hand side, and assume an interpretation $\omega = \langle c, id, d \rangle \gg\!= \tau$, consisting of a state $\tau = \omega[0, 1, 0] \in \mathcal{D}(B)$ and channels $c \colon B \to A$ and $d \colon B \to C$. We write $\sigma = c \gg\!= \tau = \omega[1, 0, 0]$ and take the Bayesian inversion $c^\dagger_\tau \colon A \to B$. We now have an interpretation of the string diagram in the middle, which is equal to

ω, since by (7.26):



Similarly one obtains an interpretation of the string diagram on the right via $\rho = d \gg\!= \tau = \omega[0, 0, 1]$ and the inversion $d_\tau^\dagger \colon C \nrightarrow B$.

In the direction ($\Leftarrow$) one uses Bayesian inversion in a similar manner to transform one interpretation into another one.

2 The direction ($\Rightarrow$) is easy and is left to the reader. In fact, Proposition 7.9.5 (1) below gives a slightly stronger result.

We concentrate on ($\Leftarrow$). By assumption, for $b \in B$ we can write:

$$(\omega|_{\mathbf{1}\otimes\mathbf{1}_b\otimes\mathbf{1}})[1, 0, 1] = \sigma_b \otimes \tau_b \quad \text{for } \sigma_b \in \mathcal{D}(A), \tau_b \in \mathcal{D}(C).$$

In a next step we define channels $f \colon B \nrightarrow A$ and $g \colon B \nrightarrow C$ and a state $\rho \in \mathcal{D}(B)$ as:

$$f(b) \coloneqq \sigma_b \qquad g(b) \coloneqq \tau_b \qquad \rho \coloneqq \omega[0, 1, 0].$$

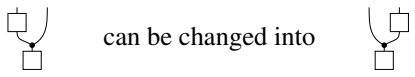
Then, for $x \in A$ and $z \in C$,

$$\begin{aligned} f(b)(x) \cdot g(b)(z) = (f(b) \otimes g(b))(x, z) &= (\sigma_b \otimes \tau_b)(x, z) \\ &= (\omega|_{\mathbf{1}\otimes\mathbf{1}_b\otimes\mathbf{1}})[1, 0, 1](x, z) \\ &= \sum\nolimits_y \frac{\omega(x, y, z) \cdot \mathbf{1}_b(y)}{\omega \models \mathbf{1} \otimes \mathbf{1}_b \otimes \mathbf{1}} \\ &= \frac{\omega(x, b, z)}{\omega[0, 1, 0] \models \mathbf{1}_b} \\ &= \frac{\omega(x, b, z)}{\rho(b)}. \end{aligned}$$

But now we see that ω has a fork shape:

$$\omega(y, b, z) = f(b)(x) \cdot g(b)(z) \cdot \rho(b) = (\langle f, id, g\rangle \gg\!= \rho)(y, b, z). \qquad \blacksquare$$

What the first item of this result shows is that (sub)shapes of the form:



and vice-versa. By applying these transformation directly to the shapes in Theorem 7.9.3 the equivalences $\Leftrightarrow$ can be obtained.

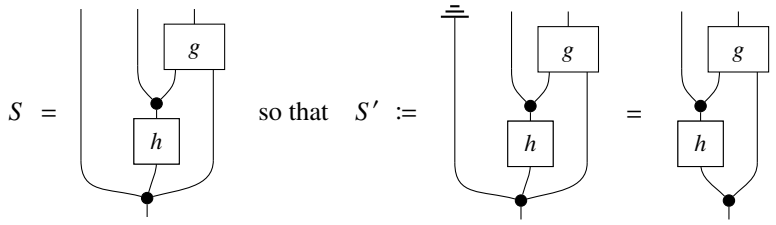
In the naive Bayesian model in Subsection 7.7.1 we have extracted certain structure from a joint state, given a particular string diagram (or shape) (7.29). This can be done quite generally, essentially as in Example 7.9.1. But note that the resulting interpretation of the string diagram need not be equal to the original joint state — as Subsection 7.7.1 shows.

Proposition 7.9.4. *Let c be a channel with full support, of the same type as an accessible string diagram $S \in SD(\Sigma)$ that does not contain $\frac{=}{\top}$. Then there is a unique interpretation of Σ that can be obtained from c. In this way we can factorise c according to S.*

In the special case that $c \models\!\approx S$ holds, the factorisation interpretation of S obtained in this way from c is the same one that gives $[\![S]\!] = c$ because of $c \models\!\approx S$ and uniqueness of disintegrations.

Proof. We conveniently write multiple wires as a single wire, using product types. We use induction on the number of boxes in S. If this number is zero, then S only consists of 'structural' elements, and can be interpreted irrespective of the channel c.

Now let S contain at least one box. Pick a box g at the top of S, with all of its output wires directly coming out of S. Thus, since S is accessible, we may assume that it has the form as described on the left below.



By discarding the wire on the left we get a diagram S' as used in disintegration, see Definition 7.5.4. But then a unique channel (interpretation) g can be extracted from $[\![S']\!] = c[0, 1, 1]$.

We can now apply the induction hypothesis with the channel $c[1, 1, 0]$ and corresponding string diagram from with the box g has been removed. It will give an interpretation of all the other boxes — not equal to g — in S. ❑

This result gives a way of testing whether a channel c has shape S: factorise c according to S and check if the resulting interpretation $[\![S]\!]$ equals c. Unfortunately, this is computationally rather expensive.

7.9.1 Shapes under conditioning

We have seen that a distribution can have a certain shape. An interesting question that arises is: what happens to such a shape when the distribution is updated? The result below answers this question, much like in [88], for three basic shapes, called fork, chain and collider,

Proposition 7.9.5. *Let $\omega \in \mathcal{D}(X \times Y \times Z)$ be an arbitrary distribution and let $q \in Fact(Y)$ be a factor on its middle component Y. We write $a \in Y$ for an arbitrary element with associated point predicate $\mathbf{1}_a$.*

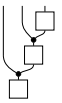
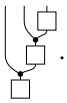
1 Let ω have fork *shape:*

$\omega \mathrel{|\!\approx}$  *then also* $\omega|_{\mathbf{1}\otimes q\otimes\mathbf{1}} \mathrel{|\!\approx}$ .

In the special case of conditioning with a point predicate we get:

$\omega|_{\mathbf{1}\otimes\mathbf{1}_a\otimes\mathbf{1}} \mathrel{|\!\approx}$.

As a consequence of Theorem 7.9.3, if ω has chain *shape*

$\omega \mathrel{|\!\approx}$  *then also* $\omega|_{\mathbf{1}\otimes q\otimes\mathbf{1}} \mathrel{|\!\approx}$ .

2 Let ω have collider *shape:*

$\omega \mathrel{|\!\approx}$  *then* $\omega|_{\mathbf{1}\otimes q\otimes\mathbf{1}} \mathrel{|\!\approx}$ 

For this shape it does not matter if q is a point predicate or not.

Proof. 1 Let's assume we have an interpretation $\omega = \langle c, id, d\rangle \gg\!= \sigma$, for $c\colon Y \nrightarrow X$, $d\colon Y \nrightarrow Z$ and $\sigma \in \mathcal{D}(Y)$. Then:

$$
\begin{aligned}
&\omega|_{\mathbf{1}\otimes q\otimes\mathbf{1}} \\
&= (\langle c, id, d\rangle \gg\!= \sigma)|_{\mathbf{1}\otimes q\otimes\mathbf{1}} \\
&= \langle c, id, d\rangle|_{\mathbf{1}\otimes q\otimes\mathbf{1}} \gg\!= (\sigma)|_{\langle c,id,d\rangle \prec\!\ll (\mathbf{1}\otimes q\otimes\mathbf{1})}) && \text{by Corrolary 6.3.12 (2)} \\
&= \langle c, id, d\rangle \gg\!= (\sigma|_q) && \text{by Exercises 6.1.11 and 4.3.8.}
\end{aligned}
$$

Hence we see the same shape that ω has.

In the special case when $q = \mathbf{1}_a$ we can extend the above calculation and

obtain a parallel product of states:

$$\begin{aligned}
\omega|_{\mathbf{1}\otimes\mathbf{1}_a\otimes\mathbf{1}} &= \langle c, id, d\rangle \gg\!= (\sigma|_{\mathbf{1}_a}) && \text{as just shown}\\
&= ((c \otimes id \otimes d) \circ \Delta_3) \gg\!= 1|a\rangle && \text{by Lemma 6.1.6 (2)}\\
&= (c \otimes id \otimes d) \gg\!= (\Delta_3 \gg\!= 1|a\rangle)\\
&= (c \otimes id \otimes d) \gg\!= (1|a\rangle \otimes 1|a\rangle \otimes 1|a\rangle)\\
&= (c \gg\!= 1|a\rangle) \otimes 1|a\rangle \otimes (d \gg\!= 1|a\rangle).
\end{aligned}$$

2 Let's assume as interpretation of the collider shape:

$$\omega \;=\; (id \otimes c \otimes id) \gg\!= ((\Delta \otimes \Delta) \gg\!= (\sigma \otimes \tau)),$$

for states $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Z)$ and channel $c\colon X \times Z \nrightarrow Y$. Then:

$$\begin{aligned}
&\omega|_{\mathbf{1}\otimes q\otimes\mathbf{1}}\\
&= \Big((id \otimes c \otimes id) \gg\!= ((\Delta \otimes \Delta) \gg\!= (\sigma \otimes \tau))\Big)\Big|_{\mathbf{1}\otimes q\otimes\mathbf{1}}\\
&= (id \otimes c|_q \otimes id) \gg\!= \Big(((\Delta \otimes \Delta) \gg\!= (\sigma \otimes \tau))\Big|_{\mathbf{1}\otimes(c \lessdot q)\otimes\mathbf{1}}\Big)\\
&\qquad \text{by Corrolary 6.3.12 (3)}\\
&= (id \otimes c|_q \otimes id) \gg\!= ((\Delta \otimes \Delta)|_{\mathbf{1}\otimes(c \lessdot q)\otimes\mathbf{1}} \gg\!= \Big((\sigma \otimes \tau)\Big|_{(\Delta\otimes\Delta) \lessdot (\mathbf{1}\otimes(c \lessdot q)\otimes\mathbf{1})}\Big)\\
&\qquad \text{by Theorem 6.3.11}\\
&= (id \otimes c|_q \otimes id) \gg\!= ((\Delta \otimes \Delta) \gg\!= \Big((\sigma \otimes \tau)\Big|_{(\Delta\otimes\Delta) \lessdot (\mathbf{1}\otimes(c \lessdot q)\otimes\mathbf{1})}\Big)\\
&\qquad \text{by Exercise 6.1.10.}
\end{aligned}$$

The updated state $(\sigma \otimes \tau)\big|_{(\Delta\otimes\Delta) \lessdot (\mathbf{1}\otimes(c \lessdot q)\otimes\mathbf{1})}$ is typically entwined, even if q is a point predicate, see also Exercise 6.1.8. ❑

The fact that conditioning with a point predicate destroys the shape is an important phenomenon since it allows us to break entwinedness / correlations. This is relevant in statistical analysis, esp. w.r.t. causality [136, 137], see the causal surgery procedure in Section **??**. In such a context, conditioning on a point predicate is often expressed in terms of 'controlling for'. For instance, if there is a gender component $G = \{m, f\}$ with elements m for male and f for female, then conditioning with a point predicate $\mathbf{1}_m$ or $\mathbf{1}_f$, suitably weakenend via tensoring with truth $\mathbf{1}$, amounts to controlling for gender. Via restriction to one gender value one fragments the shape and thus controls the influence of gender in the situation at hand.

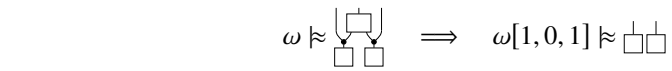
Exercises

7.9.1 Check the aim of Exercise 2.5.2 really is to prove the shape statement

$$\omega \models\!\approx$$

for the (ternary) state ω defined there.

7.9.2 Prove that a collider shape leads to non-entwinedness:



7.9.3 Let $S \in SD(\Sigma)$ be given with an interpretation of the string diagram signature Σ. Check that then $[\![S]\!] \models\!\approx S$.

7.9.4 Prove the following items, which are known as the 'semigraphoid' properties, see [163, 55]; they are seen as the basic rules of conditional independence.

1 Symmetry: $\omega \models\!\approx$  implies $\langle \pi_3, \pi_2, \pi_1 \rangle \gg\!= \omega \models\!\approx$ .

2 Decomposition: $\omega \models\!\approx$  implies $\omega[0, 1, 1, 1] \models\!\approx$ 

3 Weak union: $\omega \models\!\approx$  implies $\omega \models\!\approx$ 

Hint: Apply disintegration to the upper-left box.

4 Contraction: if $\omega \models\!\approx$  and also $\omega[1, 1, 1, 0] \models\!\approx$  then $\omega \models\!\approx$ .

Hint: Form a suitable combination of the two upper-right boxes in the assumptions.

7.10 Inference in Bayesian networks, reconsidered

Inference is one of the main topics in Bayesian probability. We have seen illustrations of Bayesian inference in Section 6.5 for the Asia Bayesian network, using forward and back inference along a channel (see Definition 6.3.1). We are now in a position to approach the topic of inference more systematically, using string diagrams. We start by defining inference itself, namely as what we have earlier called crossover inference.

Let $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n)$ be a joint state on a product space $X_1 \times \cdots \times X_n$. Suppose we have evidence p on one component X_i, in the form of a factor $p \in Fact(X_i)$. After updating with p, we are interested in the marginal distribution at component j. We will assume that $i \neq j$, otherwise the problem is simple and can be reduced to updating the marginal at $i = j$ with p, see Lemma 6.1.6 (6).

In order to update ω on $X_1 \times \cdots \times X_n$ with $p \in Fact(X_i)$ we first have to

extend the factor p to a factor on the whole product space, by weakening it to:

$$\pi_i \mathrel{=\!\!\ll} p \;=\; \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes p \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1}.$$

After the update with this factor we take the marginal at j in the form of:

$$\pi_j \gg\!= (\omega|_{\pi_i \mathrel{=\!\!\ll} p}) \;=\; \omega|_{\pi_i \mathrel{=\!\!\ll} p}[0, \ldots, 0, 1, 0 \ldots, 0].$$

The latter provides a convenient form.

Definition 7.10.1. A basic *inference query* is of the form:

$$\pi_j \gg\!= (\omega|_{\pi_i \mathrel{=\!\!\ll} p}) \tag{7.34}$$

for a joint state ω with an evidence factor p on its i-th component and with the j-th marginal as conclusion.

In words: an inference query is a marginalisation of a joint state conditioned with a weakened factor. What is called *inference* is the activity of calculating the outcome of an inference query.

The term 'query' suggests a connection to the world of databases, especially in probabilistic form [156]. This is justified since in the above expression $\pi_j \gg\!= (\omega|_{\pi_i \mathrel{=\!\!\ll} p})$ in (7.34) one may read the select-from-where structure in basic database queries:

- The marginalisation $\pi_j \gg\!= (-)$ performs the select-part, performing a restriction to the j-th component.
- The joint distribution ω on tuples can be seen as a probabilistic database.
- The conditioning $|$ with weakened predicate $\pi_i \mathrel{=\!\!\ll} p$ corresponds to the where-part that imposes a condition on the database.

An inference query may also have multiple evidence factors $p_k \in \mathit{Fact}(X_{i_k})$, giving an inference query of the form:

$$\pi_j \gg\!= (\omega|_{(\pi_{i_1} \mathrel{=\!\!\ll} p_1) \,\&\, \cdots \,\&\, (\pi_{i_m} \mathrel{=\!\!\ll} p_{i_m})}) \;=\; \pi_j \gg\!= (\omega|_{\pi_{i_1} \mathrel{=\!\!\ll} p_1} \;\cdots\; |_{\pi_{i_m} \mathrel{=\!\!\ll} p_{i_m}}).$$

By suitably swapping components and using products one can reduce such a query to one of the from (7.34). Similarly one may marginalise on several components at the same time and again reduce this to the canonical form (7.34).

The notion of inference that we use is formulated in terms of joint states. This gives mathematical clarity, but not a practical method to actually compute queries. If the joint state has the shape of a Bayesian network, we can use this network structure to guide the computations. This is formalised in the next result: it describes quite generally what we have been illustrating many times already, namely that inference can be done along channels — both forward and

backward — in particular along channels that interpret edges in a Bayesian network (*i.e.* conditional probability tables).

Theorem 7.10.2. *Let joint state ω have the shape S of a Bayesian network: $\omega \models\!\approx S$. An inference query $\pi_j \gg\!= (\omega|_{\pi_i \asymp\!\ll p})$ can then be computed via forward and backward inference along channels in the string diagram S.*

Proof. We use induction on the number of boxes in S. If this number is one, S consists of a single state box, whose interpretation is ω. Then we are done by the crossover inference Corollary 6.3.13.

We now assume that S contains more than one box. We pick a box at the top of S so that we can write:

$$\omega \;=\; (id_{X_1 \times \cdots \times X_{k-1}} \otimes c \otimes id_{X_{k+1} \times \cdots \times X_n}) \gg\!= \rho \qquad \text{where} \qquad \rho \models\!\approx S',$$

and where S' is the string diagram obtained from S by removing the single box whose interpretation we have written as channel c, say of type $Y \to\!\!\!\!\bullet\; X_k$. The induction hypothesis applies to ρ and S'.

Consider the situation below wrt. the underlying product space $X_1 \times \cdots \times X_n$ of ω,

$$\begin{array}{ccccccccccccc} & & & & \text{evidence} & & & & & & \text{marginal} & & \\ & & & & {\scriptstyle p}\downarrow & & & & & & \downarrow & & \\ X_1 & \times \cdots \times & X_i & \times \cdots \times & X_k & \times \cdots \times & X_j & \times \cdots \times & X_n \\ & & & & \uparrow{\scriptstyle c} & & & & \\ & & & & \text{channel} & & & & \end{array}$$

Below we distinguish three cases about (in)equality of i, k, j, where, recall, that we assume $i \neq j$.

1 $i \neq k$ and $j \neq k$. In that case we can compute the inference query as:

$$\begin{aligned} \pi_j \gg\!= (\omega|_{\pi_i \asymp\!\ll p}) \;&=\; \pi_j \gg\!= \Big((id \otimes c \otimes id) \gg\!= \rho)|_{\pi_i \asymp\!\ll p}\Big) \\ &=\; \pi_j \gg\!= \Big((id \otimes c \otimes id) \gg\!= (\rho|_{\pi_i \asymp\!\ll p})\Big) \quad \text{by Exercise 6.1.7 (1)} \\ &=\; \pi_j \gg\!= (\rho|_{\pi_i \asymp\!\ll p}). \end{aligned}$$

Hence we are done by the induction hypothesis.

Aside: we are oversimplifying by assuming that the domain of the channel c is not a product space; if it is, say $Y_1 \times \cdots \times Y_m$ of length $m > 1$ we should not marginalise at j but at $j + m - 1$. A similar shift may be needed for the weakening $\pi_i \asymp\!\ll p$ if $i < k$.

2 $i = k$, and hence $j \neq k$. Then:

$$\begin{aligned}\pi_j \gg\!= (\omega|_{\pi_i \,=\!\ll p}) &= \pi_j \gg\!= \Big((id \otimes c \otimes id) \gg\!= \rho)|_{\pi_i \,=\!\ll p}\Big) \\ &= \pi_j \gg\!= (\rho|_{(id\otimes c\otimes id) \,=\!\ll (\pi_i \,=\!\ll p)}) \qquad \text{by Exercise 6.1.7 (2)} \\ &= \pi_j \gg\!= (\rho|_{\pi_i \,=\!\ll (c \,=\!\ll p)}).\end{aligned}$$

The induction hypothesis now applies with transformed evidence $c \,=\!\ll p$.

Aside: again we are oversimplifying; when the channel c has a product space as domain, the single projection π_i in the conditioning-factor $\pi_i \,=\!\ll (c \,=\!\ll p)$ may have to be replaced by an appropriate tuple of projections.

3 $j = k$, and hence $i \neq k$. Then:

$$\begin{aligned}\pi_j \gg\!= (\omega|_{\pi_i \,=\!\ll p}) &= \pi_j \gg\!= \Big((id \otimes c \otimes id) \gg\!= \rho)|_{\pi_i \,=\!\ll p}\Big) \\ &= \pi_j \gg\!= \Big((id \otimes c \otimes id) \gg\!= (\rho|_{\pi_i \,=\!\ll p})\Big) \quad \text{by Exercise 6.1.7 (1)} \\ &= c \gg\!= \Big(\pi_j \gg\!= (\rho|_{\pi_i \,=\!\ll p})\Big).\end{aligned}$$

Again we are done by the induction hypothesis.

Instead of a channel c, as used above, one can have a diagonal or a swap map in the string diagram S. Since diagonals and swaps are given by functions f — forming trivial channels — Lemma 6.1.6 (7) applies, so that we can proceed via evaluation of $\pi_i \circ f$. ❑

Example 7.10.3. We shall recompute the query 'smoking, given a positive xray' in the Asia Bayesian network from Section 6.5, this time using the mechanism of Theorem 7.10.2. In order to do this we have to choose a particular order in the channels of the network. Figure 7.5 give two such orders, which we also call 'stretchings' of the network. These two string diagrams have the same semantics, so it does not matter whether which one we take. For this illustration we choose the one on the left.

This means that we can write the joint state $\omega \in \mathcal{D}(S \times D \times X)$ as composite:

$$\begin{aligned}\omega \;=\; &(id \otimes dysp \otimes id) \circ (id \otimes id \otimes id \otimes xray) \circ (id \otimes id \otimes \Delta_2) \\ &\circ (id \otimes id \otimes either) \circ (id \otimes id \otimes id \otimes tub) \circ (id \otimes id \otimes id \otimes asia) \\ &\circ (id \otimes id \otimes lung) \circ (id \otimes dysp \otimes id) \circ \Delta_3 \circ smoking\end{aligned}$$

The positive xray evidence translates into a point predicate $\mathbf{1}_x$ on the set $X = \{x, x^{\perp}\}$ of xray outcomes. It has to be weakened to $\pi_3 \,=\!\ll \mathbf{1}_x = \mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}_x$ on the product space $S \times D \times X$. We are interested in the updated smoking distribution, which can be obtained as first marginal. Thus we compute the following inference query, where the number k in $\overset{k}{=}$ refers to the k-th of the

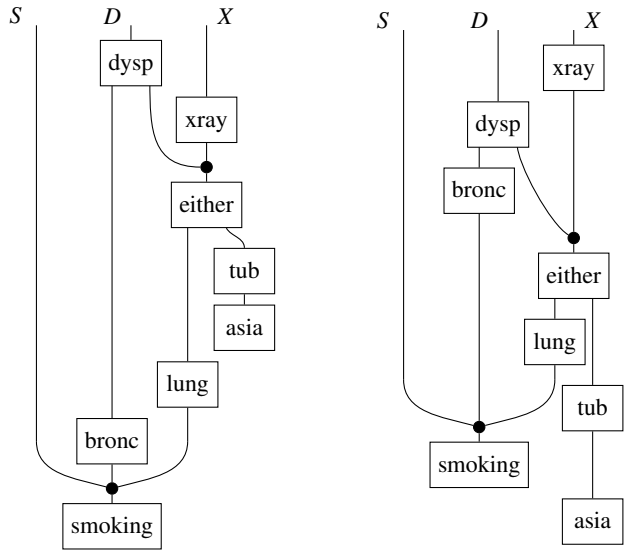


Figure 7.5 Two 'stretchings' of the Asia Bayesian network from Section 6.5, with the same semantics.

three distinctions in the proof of Theorem 7.10.2.

$$
\begin{aligned}
&\pi_1 \gg\!= \left(\omega\big|_{\pi_3 \prec\!\ll \mathbf{1}_x}\right) \\
&= \pi_1 \gg\!= \Big(((id \otimes dysp \otimes id) \circ \cdots)\big|_{\pi_3 \prec\!\ll \mathbf{1}_x}\Big) \\
&\overset{1}{=} \pi_1 \gg\!= \Big(((id \otimes id \otimes id \otimes xray) \circ \cdots)\big|_{\pi_4 \prec\!\ll \mathbf{1}_x}\Big) \\
&\overset{2}{=} \pi_1 \gg\!= \Big(((id \otimes id \otimes \Delta_2) \circ \cdots)\big|_{\pi_4 \prec\!\ll (xray \prec\!\ll \mathbf{1}_x)}\Big) \\
&= \pi_1 \gg\!= \Big(((id \otimes id \otimes either) \circ \cdots)\big|_{\pi_3 \prec\!\ll (xray \prec\!\ll \mathbf{1}_x)}\Big) \\
&\overset{2}{=} \pi_1 \gg\!= \Big(((id \otimes id \otimes id \otimes tub) \circ \cdots)\big|_{\langle \pi_3, \pi_4 \rangle \prec\!\ll (either \prec\!\ll (xray \prec\!\ll \mathbf{1}_x))}\Big) \\
&\overset{2}{=} \pi_1 \gg\!= \Big(((id \otimes id \otimes id \otimes asia) \circ \cdots)\big|_{\langle \pi_3, \pi_4 \rangle \prec\!\ll ((id \otimes tub) \prec\!\ll (either \prec\!\ll (xray \prec\!\ll \mathbf{1}_x)))}\Big) \\
&\overset{2}{=} \pi_1 \gg\!= \Big(((id \otimes id \otimes lung) \circ \cdots)\big|_{\pi_3 \prec\!\ll ((id \otimes asia) \prec\!\ll ((id \otimes tub) \prec\!\ll (either \prec\!\ll (xray \prec\!\ll \mathbf{1}_x))))}\Big) \\
&\overset{2}{=} \pi_1 \gg\!= \Big(((id \otimes dysp \otimes id) \circ \cdots)\big|_{\pi_3 \prec\!\ll (lung \prec\!\ll ((id \otimes asia) \prec\!\ll ((id \otimes tub) \prec\!\ll (either \prec\!\ll (xray \prec\!\ll \mathbf{1}_x)))))}\Big) \\
&\overset{1}{=} \pi_1 \gg\!= \Big((\Delta_3 \circ smoking)\big|_{\pi_3 \prec\!\ll (lung \prec\!\ll ((id \otimes asia) \prec\!\ll ((id \otimes tub) \prec\!\ll (either \prec\!\ll (xray \prec\!\ll \mathbf{1}_x)))))}\Big) \\
&= smoking\big|_{lung \prec\!\ll ((id \otimes asia) \prec\!\ll ((id \otimes tub) \prec\!\ll (either \prec\!\ll (xray \prec\!\ll \mathbf{1}_x))))} \\
&= smoking\big|_{(xray \circ either \circ (id \otimes tub) \circ (id \otimes asia) \circ lung) \prec\!\ll \mathbf{1}_x} \\
&= 0.6878|s\rangle + 0.3122|s^{\perp}\rangle.
\end{aligned}
$$

It is a exercise below to compute the same inference query for the stretching

on the right in Figure 7.5. The outcome is necessarily the same, by Theorem 7.10.2, since there it is shown that all such inference computations are equal to the computation on the joint state.

The inference calculation in Example 7.10.3 is quite mechanical in nature and can thus be implemented easily, giving a channel-based inference algorithm, see [75] for Bayesian networks. The algorithm consists of two parts.

1 It first finds a stretching of the Bayesian network with a minimal *width*, that is, a description of the network as a sequence of channels such that the state space in between the channels has minimal size. As can be seen in Figure 7.5, there can be quite a bit of freedom in choosing the order of channels.
2 Perform the calculation as in Example 7.10.3, following the steps in the proof of Theorem 7.10.2.

The resulting algorithm's performance compares favourably to the performance of the `pgmpy` Python library for Bayesian networks [5]. By design, it uses (fuzzy) factors and not (sharp) events and thus solves the "soft evidential update" problem [160].

Exercises

7.10.1 In Example 7.10.3 we have identified a state on a space X with a channel $1 \to X$, as we have done before, *e.g.* in Exercise 1.8.2. Consider states $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$ with a factor $p \in \mathit{Fact}(X \times Y)$. Prove that, under this identification,

$$\sigma|_{(id \otimes \tau) \mathrel{=\!\!\ll} p} = (\sigma \otimes \tau)|_p[1, 0].$$

7.10.2 Compute the inference query from Example 7.10.3 for the stretching of the Asia Bayesian network on the right in Figure 7.5.

7.10.3 From Theorem 7.10.2 one can conclude that the for the channel-based calculation of inference queries the particular stretching of a Bayesian network does not matter — since they are all equal to inference on the joint state. Implicitly, there is a 'shift' result about forward and backward inference.

Let $c \colon A \to X$ and $d \colon B \to Y$ be channels, together with a joint state $\omega \in \mathcal{D}(A \times X)$ and a factor $q \colon Y \to \mathbb{R}_{\geq 0}$ on Y. Then the following marginal distributions are the same.

$$\begin{aligned} &\Big(((c \otimes id) \gg\!= \omega)\big|_{(id \otimes d) \mathrel{=\!\!\ll} (\mathbf{1} \otimes q)}\Big)[1, 0] \\ &= \Big((c \otimes id) \gg\!= (((id \otimes d) \gg\!= \omega)\big|_{\mathbf{1} \otimes q})\Big)[1, 0]. \end{aligned}$$

1 Prove this equation.
2 Give an interpretation of this equation in relation to inference calculations.

References

[1] S. Abramsky. Contextual semantics: From quantum mechanics to logic, databases, constraints, and complexity. *EATCS Bulletin*, 113, 2014.

[2] D. Aldous. Exchangeability and related topics. In P. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, number 1117 in Lect. Notes Math., pages 1–198. Springer, Berlin, 1985.

[3] E. Alfsen and F. Shultz. *State spaces of operator algebras: basic theory, orientations and C*-products*. Mathematics: Theory & Applications. Birkhauser Boston, 2001.

[4] C. Anderson. The end of theory. The data deluge makes the scientific method obsolete. *Wired*, 16:07, 2008. Available at `https://www.wired.com/2008/06/pb-theory/`.

[5] A. Ankand and A. Panda. *Mastering Probabilistic Graphical Models using Python*. Packt Publishing, Birmingham, 2015.

[6] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1152–1174., 1974.

[7] S. Awodey. *Category Theory*. Oxford Logic Guides. Oxford Univ. Press, 2006.

[8] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge Univ. Press, 2012. Publicly available via `http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Brml.HomePage`.

[9] M. Barr and Ch. Wells. *Toposes, Triples and Theories*. Springer, Berlin, 1985. Revised and corrected version available from URL: `www.tac.mta.ca/tac/reprints/articles/12/tr12.pdf`.

[10] M. Barr and Ch. Wells. *Category Theory for Computing Science*. Prentice Hall, Englewood Cliffs, NJ, 1990.

[11] G. Barthe and J. Hsu. Probabilistic couplings from program logics. In G. Barthe, J.-P. Katoen, and A. Silva, editors, *Foundations of Probabilistic Programming*, pages 145–184. Cambridge Univ. Press, 2021.

[12] J. Beck. Distributive laws. In B. Eckman, editor, *Seminar on Triples and Categorical Homology Theory*, number 80 in Lect. Notes Math., pages 119–140. Springer, Berlin, 1969.

[13] J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley & Sons, 2000.

[14] C. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[15] F. van Breugel, C. Hermida, M. Makkai, and J. Worrell. Recursively defined metric spaces without contraction. *Theor. Comp. Sci.*, 380:143–163, 2007.

[16] P. Bruza and S. Abramsky. Probabilistic programs: Contextuality and relational database theory. In *Quantum Interaction*, pages 163–174, 2016.

[17] F. Buscemi and V. Scarani. Fluctuation theorems from bayesian retrodiction. *Phys. Rev. E*, 103:052111, 2021.

[18] A. Carboni and R. Walters. Cartesian bicategories I. *Journ. of Pure & Appl. Algebra*, 49(1-2):11–32, 1987.

[19] H. Chan and A. Darwiche. On the revision of probabilistic beliefs using uncertain evidence. *Artif. Intelligence*, 163:67–90, 2005.

[20] K. Cho and B. Jacobs. The EfProb library for probabilistic calculations. In F. Bonchi and B. König, editors, *Conference on Algebra and Coalgebra in Computer Science (CALCO 2017)*, volume 72 of *LIPIcs*. Schloss Dagstuhl, 2017.

[21] K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, 29(7):938–971, 2019.

[22] K. Cho, B. Jacobs, A. Westerbaan, and B. Westerbaan. An introduction to effectus theory. see `arxiv.org/abs/1512.05813`, 2015.

[23] A. Clark. *Surfing Uncertainty. Prediction, Action, and the Embodied Mind*. Oxford Univ. Press, 2016.

[24] F. Clerc, F. Dahlqvist, V. Danos, and I. Garnier. Pointless learning. In J. Esparza and A. Murawski, editors, *Foundations of Software Science and Computation Structures*, number 10203 in Lect. Notes Comp. Sci., pages 355–369. Springer, Berlin, 2017.

[25] B. Coecke and A. Kissinger. *Picturing Quantum Processes. A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge Univ. Press, 2016.

[26] B. Coecke and R. Spekkens. Picturing classical and quantum Bayesian inference. *Synthese*, 186(3):651–696, 2012.

[27] H. Crane. The ubiquitous Ewens sampling formula. *Statistical Science*, 31(1):1–19, 2016.

[28] J. Culbertson and K. Sturtz. A categorical foundation for Bayesian probability. *Appl. Categorical Struct.*, 22(4):647–662, 2014.

[29] F. Dahlqvist, V. Danos, and I. Garnier. Robustly parameterised higher-order probabilistic models. In J. Desharnais and R. Jagadeesan, editors, *Int. Conf. on Concurrency Theory*, volume 59 of *LIPIcs*, pages 23:1–23:15. Schloss Dagstuhl, 2016.

[30] F. Dahlqvist and D. Kozen. Semantics of higher-order probabilistic programs with conditioning. In *Princ. of Programming Languages*, pages 57:1–57:29. ACM Press, 2020.

[31] F. Dahlqvist, L. Parlant, and A. Silva. Layer by layer – combining monads. In B. Fischer and T. Uustalu, editors, *Theoretical Aspects of Computing*, number 11187 in Lect. Notes Comp. Sci., pages 153–172. Springer, Berlin, 2018.

[32] V. Danos and T. Ehrhard. Probabilistic coherence spaces as a model of higher-order probabilistic computation. *Information & Computation*, 209(6):966–991, 2011.

[33] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge Univ. Press, 2009.

[34] S. Dash and S. Staton. A monad for probabilistic point processes. In D. Spivak and J. Vicary, editors, *Applied Category Theory Conference*, Elect. Proc. in Theor. Comp. Sci., 2020.

[35] S. Dash and S. Staton. Monads for measurable queries in probabilistic databases. In A. Sokolova, editor, *Math. Found. of Programming Semantics*, 2021.

[36] E. Davies and J. Lewis. An operational approach to quantum probability. *Communic. Math. Physics*, 17:239–260, 1970.

[37] B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Memorie della R. Accademia Nazionale dei Lincei*, IV, fasc. 5:86–113, 1930. Available at `www.brunodefinetti.it/Opere/funzioneCaratteristica.pdf`.

[38] B. de Finetti. *Theory of Probability: A critical introductory treatment*. Wiley, 2017.

[39] P. Diaconis and S. Zabell. Updating subjective probability. *Journ. American Statistical Assoc.*, 77:822–830, 1982.

[40] P. Diaconis and S. Zabell. Some alternatives to Bayes' rule. Technical Report 339, Stanford Univ., Dept. of Statistics, 1983.

[41] F. Dietrich, C. List, and R. Bradley. Belief revision generalized: A joint characterization of Bayes' and Jeffrey's rules. *Journ. of Economic Theory*, 162:352–371, 2016.

[42] E. Dijkstra. *A Discipline of Programming*. Prentice Hall, Englewood Cliffs, NJ, 1976.

[43] E. Dijkstra and C. Scholten. *Predicate Calculus and Program Semantics*. Springer, Berlin, 1990.

[44] A. Dvurečenskij and S. Pulmannová. *New Trends in Quantum Structures*. Kluwer Acad. Publ., Dordrecht, 2000.

[45] M. Erwig and E. Walkingshaw. A DSL for explaining probabilistic reasoning. In W. Taha, editor, *Domain-Specific Languages*, number 5658 in Lect. Notes Comp. Sci., pages 335–359. Springer, Berlin, 2009.

[46] W. Ewens. The sampling theory of selectively neutral alleles. *Theoret. Population Biology*, 3:87–112, 1972.

[47] W. Feller. *An Introduction to Probability Theory and Its applications*, volume I. Wiley, 3rd rev. edition, 1970.

[48] B. Fong. Causal theories: A categorical perspective on Bayesian networks. Master's thesis, Univ. of Oxford, 2012. see `arxiv.org/abs/1301.6201`.

[49] D. J. Foulis and M.K. Bennett. Effect algebras and unsharp quantum logics. *Found. Physics*, 24(10):1331–1352, 1994.

[50] S. Friedland and S. Karlin. Some inequalities for the spectral radius of nonnegative matrices and applications. *Duke Math. Journ.*, 42(3):459–490, 1975.

[51] K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[52] T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, 370:107239, 2020.

[53] T. Fritz and E. Rischel. Infinite products and zero-one laws in categorical probability. *Compositionality*, 2(3), 2020.

[54] S. Fujii, S. Katsumata, and P. Melliès. Towards a formal theory of graded monads. In B. Jacobs and C. Löding, editors, *Foundations of Software Science and*

Computation Structures, number 9634 in Lect. Notes Comp. Sci., pages 513–530. Springer, Berlin, 2016.

[55] D. Geiger, T., and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20:507–534, 1990.

[56] A. Gibbs and F. Su. On choosing and bounding probability metrics. *Int. Statistical Review*, 70(3):419–435, 2002.

[57] G. Gigerenzer and U. Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704, 1995.

[58] M. Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, number 915 in Lect. Notes Math., pages 68–85. Springer, Berlin, 1982.

[59] A. Gordon, T. Henzinger, A. Nori, and S. Rajamani. Probabilistic programming. In *Int. Conf. on Software Engineering*, 2014.

[60] T. Griffiths, C. Kemp, and J. Tenenbaum. Bayesian models of cognition. In R. Sun, editor, *Cambridge Handbook of Computational Cognitive Modeling*, pages 59–100. Cambridge Univ. Press, 2008.

[61] J. Halpern. *Reasoning about Uncertainty*. MIT Press, Cambridge, MA, 2003.

[62] M. Hayhoe, F. Alajaji, and B. Gharesifard. A pólya urn-based model for epidemics on networks. In *American Control Conference*, pages 358–363, 2017.

[63] W. Hino, H. Kobayashi I. Hasuo, and B. Jacobs. Healthiness from duality. In *Logic in Computer Science*. IEEE, Computer Science Press, 2016.

[64] J. Hohwy. *The Predictive Mind*. Oxford Univ. Press, 2013.

[65] F. Hoppe. Pólya-like urns and the Ewens' sampling formula. *Journ. Math. Biology*, 20:91–94, 1984.

[66] M. Hyland and J. Power. The category theoretic understanding of universal algebra: Lawvere theories and monads. In L. Cardelli, M. Fiore, and G. Winskel, editors, *Computation, Meaning, and Logic: Articles dedicated to Gordon Plotkin*, number 172 in Elect. Notes in Theor. Comp. Sci., pages 437–458. Elsevier, Amsterdam, 2007.

[67] B. Jacobs. Convexity, duality, and effects. In C. Calude and V. Sassone, editors, *IFIP Theoretical Computer Science 2010*, number 82(1) in IFIP Adv. in Inf. and Comm. Techn., pages 1–19. Springer, Boston, 2010.

[68] B. Jacobs. New directions in categorical logic, for classical, probabilistic and quantum logic. *Logical Methods in Comp. Sci.*, 11(3), 2015.

[69] B. Jacobs. Affine monads and side-effect-freeness. In I. Hasuo, editor, *Coalgebraic Methods in Computer Science (CMCS 2016)*, number 9608 in Lect. Notes Comp. Sci., pages 53–72. Springer, Berlin, 2016.

[70] B. Jacobs. *Introduction to Coalgebra. Towards Mathematics of States and Observations*. Number 59 in Tracts in Theor. Comp. Sci. Cambridge Univ. Press, 2016.

[71] B. Jacobs. Introduction to coalgebra. Towards mathematics of states and observations. Cambridge Univ. Press, to appear, 2016.

[72] B. Jacobs. Hyper normalisation and conditioning for discrete probability distributions. *Logical Methods in Comp. Sci.*, 13(3:17), 2017. See `https://lmcs.episciences.org/3885`.

[73] B. Jacobs. A note on distances between probabilistic and quantum distributions. In A. Silva, editor, *Math. Found. of Programming Semantics*, Elect. Notes in Theor. Comp. Sci. Elsevier, Amsterdam, 2017.

[74] B. Jacobs. A recipe for state and effect triangles. *Logical Methods in Comp. Sci.*, 13(2), 2017. See `https://lmcs.episciences.org/3660`.

[75] B. Jacobs. A channel-based exact inference algorithm for Bayesian networks. See `arxiv.org/abs/1804.08032`, 2018.

[76] B. Jacobs. Learning along a channel: the Expectation part of Expectation-Maximisation. In B. König, editor, *Math. Found. of Programming Semantics*, number 347 in Elect. Notes in Theor. Comp. Sci., pages 143–160. Elsevier, Amsterdam, 2019.

[77] B. Jacobs. The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule. *Journ. of Artif. Intelligence Research*, 65:783–806, 2019.

[78] B. Jacobs. From multisets over distributions to distributions over multisets. In *Logic in Computer Science*. IEEE, Computer Science Press, 2021.

[79] B. Jacobs. Learning from what's right and learning from what's wrong. In A. Sokolova, editor, *Math. Found. of Programming Semantics*, 2021.

[80] B. Jacobs. Multinomial and hypergeometric distributions in Markov categories. In A. Sokolova, editor, *Math. Found. of Programming Semantics*, 2021.

[81] B. Jacobs. Multisets and distributions, in drawing and learning. In A. Palmigiano and M. Sadrzadeh, editors, *Samson Abramsky on Logic and Structure in Computer Science and Beyond*. Springer, 2021, to appear.

[82] B. Jacobs, A. Kissinger, and F. Zanasi. Causal inference by string diagram surgery. In M. Bojańczyk and A. Simpson, editors, *Foundations of Software Science and Computation Structures*, number 11425 in Lect. Notes Comp. Sci., pages 313–329. Springer, Berlin, 2019.

[83] B. Jacobs, J. Mandemaker, and R. Furber. The expectation monad in quantum foundations. *Information & Computation*, 250:87–114, 2016.

[84] B. Jacobs, A. Silva, and A. Sokolova. Trace semantics via determinization. *Journ. of Computer and System Sci.*, 81(5):859–879, 2015.

[85] B. Jacobs and S. Staton. De Finetti's construction as a categorical limit. In D. Petrişan and J. Rot, editors, *Coalgebraic Methods in Computer Science (CMCS 2020)*, number 12094 in Lect. Notes Comp. Sci., pages 90–111. Springer, Berlin, 2020.

[86] B. Jacobs and A. Westerbaan. Distances between states and between predicates. *Logical Methods in Comp. Sci.*, 16(1), 2020. See `https://lmcs.episciences.org/6154`.

[87] B. Jacobs and F. Zanasi. A predicate/state transformer semantics for Bayesian learning. In L. Birkedal, editor, *Math. Found. of Programming Semantics*, number 325 in Elect. Notes in Theor. Comp. Sci., pages 185–200. Elsevier, Amsterdam, 2016.

[88] B. Jacobs and F. Zanasi. A formal semantics of influence in Bayesian reasoning. In K. Larsen, H. Bodlaender, and J.-F. Raskin, editors, *Math. Found. of Computer Science*, volume 83 of *LIPIcs*, pages 21:1–21:14. Schloss Dagstuhl, 2017.

[89] B. Jacobs and F. Zanasi. The logical essentials of Bayesian reasoning. In G. Barthe, J.-P. Katoen, and A. Silva, editors, *Foundations of Probabilistic Programming*, pages 295–331. Cambridge Univ. Press, 2021.

[90] R. Jeffrey. *The Logic of Decision*. The Univ. of Chicago Press, 2nd rev. edition, 1983.

[91] F. Jensen and T. Nielsen. *Bayesian Networks and Decision Graphs*. Statistics for Engineering and Information Science. Springer, 2nd rev. edition, 2007.

[92] P. Johnstone. *Stone Spaces*. Number 3 in Cambridge Studies in Advanced Mathematics. Cambridge Univ. Press, 1982.

[93] C. Jones. *Probabilistic Non-determinism*. PhD thesis, Edinburgh Univ., 1989.

[94] P. Joyce. Partition structures and sufficient statistics. *Journ. of Applied Probability*, 35(3):622–632, 1998.

[95] A. Jung and R. Tix. The troublesome probabilistic powerdomain. In A. Edalat, A. Jung, K. Keimel, and M. Kwiatkowska, editors, *Comprox III, Third Workshop on Computation and Approximation*, number 13 in Elect. Notes in Theor. Comp. Sci., pages 70–91. Elsevier, Amsterdam, 1998.

[96] D. Jurafsky and J. Martin. Speech and language processing. Third Edition draft, available at `https://web.stanford.edu/~jurafsky/slp3/`, 2018.

[97] E. Kamenica and M. Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[98] S. Karlin. *Mathematical Methods and Theory in Games, Programming, and Economics. Vol I: Martrix Games, Programming, and Mathematical Economics*. The Java Series. Addison-Wesley, 1959.

[99] K. Keimel and G. Plotkin. Mixed powerdomains for probability and nondeterminism. *Logical Methods in Comp. Sci.*, 13(1), 2017. See `https://lmcs.episciences.org/2665`.

[100] J. Kingman. Random partitions in population genetics. *Proc. Royal Soc.*, Series A, 361:1–20, 1978.

[101] J. Kingman. The representation of partition structures. *Journ. London Math. Soc.*, 18(2):374–380, 1978.

[102] A. Kock. Closed categories generated by commutative monads. *Journ. Austr. Math. Soc.*, XII:405–424, 1971.

[103] A. Kock. Monads for which structures are adjoint to units. *Journ. of Pure & Appl. Algebra*, 104:41–59, 1995.

[104] A. Kock. Commutative monads as a theory of distributions. *Theory and Appl. of Categories*, 26(4):97–131, 2012.

[105] D. Koller and N. Friedman. *Probabilistic Graphical Models. Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

[106] D. Kozen. Semantics of probabilistic programs. *Journ. Comp. Syst. Sci*, 22(3):328–350, 1981.

[107] D. Kozen. A probabilistic PDL. *Journ. Comp. Syst. Sci*, 30(2):162–178, 1985.

[108] P. Lau, T. Koo, and C. Wu. Spatial distribution of tourism activities: A Pólya urn process model of rank-size distribution. *Journ. of Travel Research*, 59(2):231–246, 2020.

[109] S. Lauritzen. *Graphical models*. Oxford Univ. Press, Oxford, 1996.

[110] S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journ. Royal Statistical Soc.*, 50(2):157–224, 1988.

[111] F. Lawvere. The category of probabilistic mappings. Unpublished manuscript, see `ncatlab.org/nlab/files/lawvereprobability1962.pdf`, 1962.

[112] P. Lax. *Linear Algebra and Its Applications*. John Wiley & Sons, 2nd edition, 2007.

[113] T. Leinster. *Basic Category Theory*. Cambridge Studies in Advanced Mathematics. Cambridge Univ. Press, 2014. Available online via `arxiv.org/abs/1612.09375`.

[114] L. Libkin and L. Wong. Some properties of query languages for bags. In C. Beeri, A. Ohori, and D. Shasha, editors, *Database Programming Languages*, Workshops in Computing, pages 97–114. Springer, Berlin, 1993.

[115] H. Mahmoud. *Pólya Urn Models*. Chapman and Hall, 2008.

[116] E. Manes. *Algebraic Theories*. Springer, Berlin, 1974.

[117] R. Mardare, P. Panangaden, and G. Plotkin. Quantitative algebraic reasoning. In *Logic in Computer Science*. IEEE, Computer Science Press, 2016.

[118] S. Mac Lane. *Categories for the Working Mathematician*. Springer, Berlin, 1971.

[119] S. Mac Lane. *Mathematics: Form and Function*. Springer, Berlin, 1986.

[120] A. McIver and C. Morgan. *Abstraction, refinement and proof for probabilistic systems*. Monographs in Comp. Sci. Springer, 2004.

[121] A. McIver, C. Morgan, and T. Rabehaja. Abstract hidden Markov models: A monadic account of quantitative information flow. In *Logic in Computer Science*, pages 597–608. IEEE, Computer Science Press, 2015.

[122] A. McIver, C. Morgan, G. Smith, B. Espinoza, and L. Meinicke. Abstract channels and their robust information-leakage ordering. In M. Abadi and S. Kremer, editors, *Princ. of Security and Trust*, number 8414 in Lect. Notes Comp. Sci., pages 83–102. Springer, Berlin, 2014.

[123] S. Milius, D. Pattinson, and L. Schröder. Generic trace semantics and graded monads. In L. Moss and P. Sobocinski, editors, *Conference on Algebra and Coalgebra in Computer Science (CALCO 2015)*, volume 35 of *LIPIcs*, pages 253–269. Schloss Dagstuhl, 2015.

[124] H. Minc. *Nonnegative Matrices*. John Wiley & Sons, 1998.

[125] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[126] E. Moggi. Notions of computation and monads. *Information & Computation*, 93(1):55–92, 1991.

[127] R. Nagel. Order unit and base norm spaces. In A. Hartkämper and H. Neumann, editors, *Foundations of Quantum Mechanics and Ordered Linear Spaces*, number 29 in Lect. Notes Physics, pages 23–29. Springer, Berlin, 1974.

[128] M. Nielsen and I. Chuang. *Quantum Computation and Quantum Information*. Cambridge Univ. Press, 2000.

[129] F. Olmedo, F. Gretz, B. Lucien Kaminski, J-P. Katoen, and A. McIver. Conditioning in probabilistic programming. *ACM Trans. on Prog. Lang. & Syst.*, 40(1):4:1–4:50, 2018.

[130] M. Ozawa. Quantum measuring processes of continuous observables. *Journ. Math. Physics*, 25:79–87, 1984.

[131] P. Panangaden. The category of Markov kernels. In C. Baier, M. Huth, M. Kwiatkowska, and M. Ryan, editors, *Workshop on Probabilistic Methods in Verification*, number 21 in Elect. Notes in Theor. Comp. Sci., pages 171–187. Elsevier, Amsterdam, 1998.

[132] V. Paulsen and M. Tomforde. Vector spaces with an order unit. *Indiana Univ. Math. Journ.*, 58-3:1319–1359, 2009.

[133] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Graduate Texts in Mathematics 118. Morgan Kaufmann, 1988.

[134] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In R. Brachman, H. Levesque, and R. Reiter, editors, *First Intl. Conf. on Principles of Knowledge Representation and Reasoning*, pages 505–516. Morgan Kaufmann, 1989.

[135] J. Pearl. Jeffrey's rule, passage of experience, and neo-Bayesianism. In Jr. H. Kyburg, editor, *Knowledge Representation and Defeasible Reasoning*, pages 245–265. Kluwer Acad. Publishers, 1990.

[136] J. Pearl. *Causality. Models, Reasoning, and Inference*. Cambridge Univ. Press, 2nd ed. edition, 2009.

[137] J. Pearl and D. Mackenzie. *The Book of Why*. Penguin Books, 2019.

[138] B. Pierce. *Basic Category Theory for Computer Scientists*. MIT Press, Cambridge, MA, 1991.

[139] H. Pishro-Nik. *Introduction to probability, statistics, and random processes*. Kappa Research LLC, 2014. Available at `https://www.probabilitycourse.com`.

[140] D. Poole and A. Mackworth. *Artificial Intelligence. Foundations of Computational Agents*. Cambridge Univ. Press, 2nd edition, 2017. Publicly available via `https://www.cs.ubc.ca/~poole/aibook/2e/html/ArtInt2e.html`.

[141] S. Pulmannová and S. Gudder. Representation theorem for convex effect algebras. *Commentationes Mathematicae Universitatis Carolinae*, 39(4):645–659, 1998.

[142] J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[143] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2:79–87, 1999.

[144] S. Ross. *Introduction to Probability Models*. Academic Press, 9th edition, 2007.

[145] S. Ross. *A first course in probability*. Pearson Education, 10th edition, 2018.

[146] S. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ, 2003.

[147] A. Ścibior, O. Kammar, M. Vákár, S. Staton, H. Yang, Y. Cai, K. Ostermann, S. Moss, C. Heunen, and Z. Ghahramani. Denotational validation of higher-order Bayesian inference. In *Princ. of Programming Languages*, pages 60:1–60:29. ACM Press, 2018.

[148] P. Scozzafava. Uniform distribution and sum modulo *m* of independent random variables. *Statistics & Probability Letters*, 18(4):313–314, 1993.

[149] P. Selinger. A survey of graphical languages for monoidal categories. In B. Coecke, editor, *New Structures in Physics*, number 813 in Lect. Notes Physics, pages 289–355. Springer, Berlin, 2011.

[150] S. Selvin. A problem in probability (letter to the editor). *Amer. Statistician*, 29(1):67, 1975.

[151] S. Sloman. *Causal Models. How People Think about the World and Its Alternatives*. Oxford Univ. Press, 2005.

[152] A. Sokolova. Probabilistic systems coalgebraically: A survey. *Theor. Comp. Sci.*, 412(38):5095–5110, 2011.

[153] S. Staton. Probabilistic programs as measures. In G. Barthe, J.-P. Katoen, and A. Silva, editors, *Foundations of Probabilistic Programming*, pages 43–74. Cambridge Univ. Press, 2021.

[154] S. Staton, H. Yang, C. Heunen, O. Kammar, and F. Wood. Semantics for probabilistic programming: higher-order functions, continuous distributions, and soft constraints. In *Logic in Computer Science*. IEEE, Computer Science Press, 2016.

[155] M. Stone. Postulates for the barycentric calculus. *Ann. Math.*, 29:25–30, 1949.

[156] D. Suciu, D. Olteanu, C. Ré, and C. Koch. *Probabilistic Databases*. Morgan and Claypool, 2011.

[157] Y. Suhov and M. Kelbert. *Probability and Statistics by Example: Volume 1, Basic Probability and Statistics*. Cambridge Univ. Press, 2005.

[158] H. Tijms. *Understanding Probability: Chance Rules in Everyday Life*. Cambridge Univ. Press, 2nd edition, 2007.

[159] A. Tversky and D. Kahneman. Evidential impact of base rates. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgement under uncertainty: Heuristics and biases*, pages 153–160. Cambridge Univ. Press, 1982.

[160] M. Valtorta, Y.-G. Kim, and J. Vomlel. Soft evidential update for probabilistic multiagent systems. *Int. Journ. of Approximate Reasoning*, 29(1):71–106, 2002.

[161] D. Varacca. *Probability, Nondeterminism and Concurrency: Two Denotational Models for Probabilistic Computation*. PhD thesis, Univ. Aarhus, 2003. BRICS Dissertation Series, DS-03-14.

[162] D. Varacca and G. Winskel. Distributing probability over non-determinism. *Math. Struct. in Comp. Sci.*, 16:87–113, 2006.

[163] T. Verma and J. Pearl. Causal networks: semantics and expressiveness. In R. Shachter, T. Levitt, L. Kanal, and J. Lemmer, editors, *Uncertainty in Artif. Intelligence*, pages 69–78. North-Holland, 1988.

[164] C. Villani. *Optimal Transport — Old and New*. Springer, Berlin Heidelberg, 2009.

[165] I. Witten, E. Frank, and M. Hall. *Data Mining – Practical Machine Learning Tools and Techniques*. Elsevier, Amsterdam, 2011.