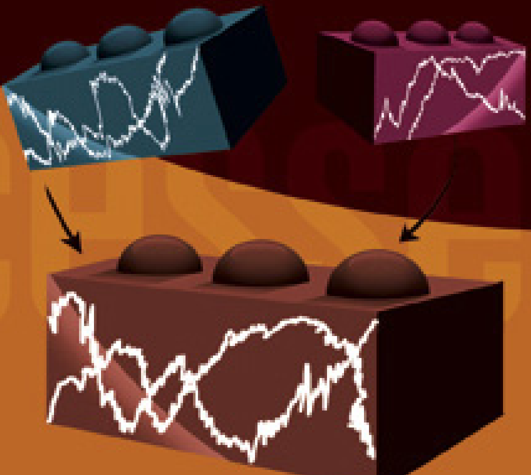
Labelled Markov Processes



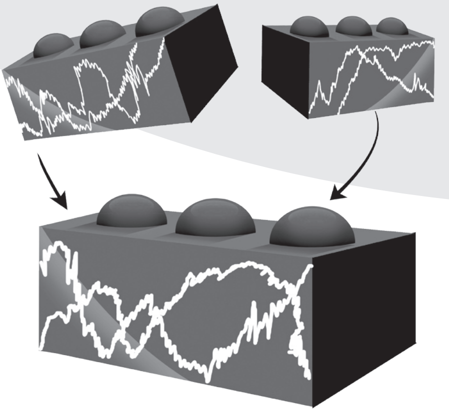
Prakash Panangaden

Imperial College Press

Labelled Markov Processes

This page intentionally left blank

Labelled Markov Processes



Prakash Panangaden

McGill University, Canada

ICP

Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Cover picture reproduced by courtesy of Jane Panangaden.

LABELLED MARKOV PROCESSES

ISBN-13 978-1-84816-287-7
ISBN-10 1-84816-287-1

Printed in Singapore.

For my father, John Anthony Panangaden, and the memory of my mother, Mary Atokaren.

This page intentionally left blank

Preface

This short book consists of two parts. The first part is an exposition of basic measure theory and probability theory on continuous state spaces. It does not replace the many excellent texts on the subject but it gives a condensed and, I hope, still readable account suitable for graduate students in computer science with a background in concurrency and semantics and without the standard undergraduate analysis curriculum. I learned this subject when I began my research into labelled Markov processes and this part of the book represents what I found useful at that time.

It is based in part on lectures that I gave at Aarhus University in 1997 and at a summer school in Udine later that same year. I had intended to let them serve as stand alone lectures but after giving a version of these lectures at Oxford in 2003–04, I was persuaded to turn them – together with an account of recent research – into book form. In the end I felt it hopeless to keep pace with the vast outpouring of papers in the last few years and have described only the work with which I was directly involved. I have, in the last chapter, given some pointers to the work of others. I feel that a comprehensive account of the whole field of probabilistic verification and related topics such as approximation and connections to planning, machine learning and performance evaluation would be premature at this point. Perhaps in the future, some more ambitious person will attempt it when the field has stabilised more.

I would like to thank Mogens Nielsen for encouraging me to teach a course on this topic at the University of Aarhus during my sabbatical year 1996-97 and Furio Honsell, Eugenio Moggi and Catuscia Palamidessi for inviting me to lecture on this at a summer school in Udine in 1997. I have benefitted from interactions with several friends, colleagues and students (those are not disjoint sets): Samson Abramsky, Christel Baier, Rick Blute, Pablo Castro, Alexandre Bouchard-Côte, Philippe Chaput, Vincent Danos, Luca de Alfaro, Erik de Vink, Luc Devroye, Ernst-Erich Doberkat, Devdatt Dubashi, Martìn Escardo, Abbas Edalat, Norm Ferns, Vineet Gupta, Holger Hermanns, Jørgen Hoffmann-Jørgensen, Chris Hundt, Joost-Pieter Katoen, Dexter Kozen, Marta Kwiatkowska, François Laviolette, Keye Martin, Mike Mislove, Gethin Norman, Catuscia Palamidessi, Joelle Pineau, Gordon Plotkin, Doina Precup, Jan Rutten, Roberto Segala, Alex Simpson, Jonathan Taylor, Franck van Breugel, James Worrell but most of all Josée Desharnais and Radha Jagadeesan. I would like to thank Jane Panangaden for the cover drawing.

Prakash Panangaden
Montreal West
September 2008

Contents

Chapter 1

Introduction

1.1 Preliminary Remarks

Labelled Markov processes (LMPs) [BDEP97; DEP02] have emerged as an important model in a variety of fields: notably artificial intelligence, verification and optimisation. The basic idea is to consider processes that exhibit stochastic behaviour but can also interact with the environment and are thus subject to nondeterminism as well. The word “labelled” is meant to suggest the process algebra notion of interaction with the environment through synchronisation on labels. These models have appeared under many other names: Markov decision processes (MDPs) [Put94], interactive Markov chains [Her02], concurrent Markov chains [Var85], probabilistic process algebras [LS91; JL91], and so on. Perhaps the most widely used term is “Markov decision process” but I use the term “labelled Markov process” to emphasise that I am not talking about rewards and the concomitant interest in policies, value functions and optimal policies.

A labelled Markov process can be thought of as a device equipped with buttons, each button with a unique label. One (or the environment) can attempt to press the buttons, the system may or may not accept the action. If it does then it makes a transition to a new state with some probability distribution on the final states. Actually, given that the action may be rejected we will use subprobability distributions. We assume that one cannot see the states, all that one can see is how the system reacts to attempts to press the buttons. This is exactly what is done in process algebra. There are variations that one can consider, for example one can have states that are partially observable; one gets hidden Markov models (HMMs) or partially observable MDPs (POMDPs). In the present book the nondeterminism appears through the presence of the labels that describe the actions of the environment, *there is no additional nondeterminism*[1].

A central concept in this book will be *bisimulation.* This concept – for purely nondeterministic systems – is due to Milner and Park but the closely related notion of lumpability in queuing theory goes back earlier and one can even argue that the roots of the idea are found in Cantor's writings. The probabilistic analogue was developed by Larsen and Skou and that work serves as the foundation for the present investigations. The main departure from their framework is the extension to systems with *continuous state spaces.* This entails mathematics that is not familiar to researchers in process algebra, programming languages and related fields. Thus one of the goals of this book is to provide the necessary background. The main application areas are to verification of stochastic hybrid systems and also to robotics and machine learning.

The theory described in this book is largely from a series of papers by the author in collaboration with Josée Desharnais, Vincent Danos, Abbas Edalat, Norm Ferns, Vineet Gupta, Radha Jagadeesan and Doina Precup as well as a few others. There have been substantial contributions by Franck van Breugel, James Worrell and their collaborators. The main contributions of these papers have been (1) a theory of bisimulation and the logical characterisation of bisimulation (2) an approximation theory for continuous-state space systems and (3) metrics for LMPs and MDPs. There have been other important contributions to the general subject of reasoning about probabilistic systems by, among others: E.-E. Doberkat [Dob03], Stoelinga and Vandraager [SV03], de Alfaro [dA97], Baier, Hermanns, Haverkort and Katoen [BHHK00], Kwiatkowska [KNP04] Segala and Lynch [SL94]; but I will not attempt to survey them. In any case the subject is still developing rapidly and any attempt to sample the current papers would go quickly out of date. Of the papers cited, nearly all work with discrete systems; de Vink and Rutten [dVR99] is one early work that explicitly attacked the problem of continuous state spaces from a coalgebraic point of view. There has been a flood of papers by Doberkat which use very sophisticated mathematics. I will not attempt to cover these either though they are of great interest.

My main goal is to make the mathematical background accessible and to give a readable self-contained account of the papers alluded to. Some of the proofs originally given have been streamlined or finessed and it is possible to present the theory in a more accessible fashion. The relevant mathematics can, of course, be learned from the many excellent standard

[1] When we discuss weak bisimulation we will allow so-called internal nondeterminism.

textbooks. However, these are thick, daunting tomes with a lot more material than strictly necessary. I have therefore chosen to write four chapters – on measure theory, integration, probability theory on continuous state spaces and on the Radon-Nikodym theorem – of essentially standard material for the benefit of the reader who has not had the opportunity to learn the material from standard sources. There is no claim to originality for this material. The chapter on the category of stochastic relations is a reworking of some fundamental ideas of Dexter Kozen [Koz81] and of Michelle Giry [Gir81]. The rest of the chapters are from the series of papers by myself and my collaborators mentioned above. We close this chapter with a review of elementary probability theory.

1.2 Elementary Discrete Probability Theory

Elementary probability theory can be summed up easily. Imagine that one has a process which makes a single step and can end up in any one of a finite set S of final states, each with equal likelihood. Then the probability that the final state lies in a subset A – often called an **event** – is given by $|A|/|S|$ where $|\cdot|$ denotes the size of a finite set. From this simple intuition one can define concepts like the probability of more complex processes which might involve several steps or interaction between different observations or situations where the outcomes are not equally likely.

The fundamental concept is that of a probability distribution.

Definition 1.1 A probability distribution on a set S is a function $P : S \longrightarrow [0, 1]$ such that $\sum_{s\in S} P(s) = 1$.

The idea is that the set S represents the possible outcomes of a random process and the number $P(s)$ is the fraction of times repeated trials of the random process is expected to yield s. This fraction may be just an estimate based on some model of the process or indeed a hunch or it may be based on statistics collected from previous trials.

Definition 1.2 A **finite probability space** is a finite set S together with a probability distribution P on S. The set S is called the **sample space**.

One can assign probabilities to sets of possible outcomes by the rule:

$$P(A \subseteq S) = \sum_{a\in A} P(a).$$

Subsets of the probability space are called *events*. The preceding formula thus extends probabilities from individual outcomes to events. It may be that one does not know, or cannot observe, the outcome completely. In this case the probability of larger events may be the best that one can do.

It may well be the case that one does not actually see the outcome of the experiment in the sense of knowing exactly the value of s at the end of the random process. More often we see some function of S.

Definition 1.3 A **random variable** on a probability space (S, P) is a function $X : S \longrightarrow T$, where T is some other set.

Most commonly we take T to be the reals $\mathbf{R}$ or perhaps the nonnegative reals $\mathbf{R}^{\geq 0}$. A random variable induces a new probability distribution on T by composition: $P_X(t \in T) := P(X^{-1}(\{t\}))$. Thus (T, P_X) becomes a new probability space.

While this notation makes sense, probabilists prefer a more set-theoretic notation which is a powerful aid to intuition. Instead of writing $X^{-1}(\{t\})$ for the set $\{s \in S | X(s) = t\}$, one writes $\{X = t\}$. We will use both notations but will prefer the latter notation whenever we are talking about probabilities and random variables, and the former when we are closer to real analysis.

When random variables take values in the reals (or in a structure where the arithmetic makes sense) we can define expectation values.

Definition 1.4 The **expectation value** of a real-valued random variable $X : S \longrightarrow \mathbf{R}$ defined on the probability space (X, P) is

$$\mathsf{E}[X] := \frac{1}{|S|} \sum_{s \in S} X(s) P(s).$$

A key notion in probability is *independence*.

Definition 1.5 Given a probability space (S, P), two events $A, B \subseteq S$ are said to be **independent** if $P(A \cap B) = P(A) \cdot P(B)$.

This is the simplest of the various independence notions. In order to appreciate independence more we need to think about how partial knowledge of the outcome of a trial affects one's estimates of other aspects of the trial.

Consider rolling two fair dice. Unless something unusual is happening we usually think of each die as being independent of the other. Thus, the probability of getting a pair of sixes, for example, is the product of the probabilities of each die showing a six which gives $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$. What if we know the sum of the values and are trying to guess the difference? Clearly if the sum is 12 the difference is 0, but if the sum is 8 there are three possible values for the difference, each with a different probability. In order to capture the general principles behind this kind of probabilistic reasoning we need the notion of *conditional probability*.

Suppose we know the outcome lies in the set B and we want to estimate whether it also lies in the set A: or what is the probability of the event A *given that* B has occurred? The original sample space is S but the knowledge that we are given cuts this down to B. Thus we have to intersect everything with B. The required probability is $P(A \cap B)/P(B)$. We assume that $P(B) \neq 0$; it would make no sense to condition on impossible events. However, later on, when we do probability on continuous state spaces, this will no longer be true and we will need to make sense of conditional probabilities more generally.

Definition 1.6 The **conditional probability** of A given B, written $P(A|B)$, is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

assuming that $P(B) \neq 0$.

It is common to use logical notation and think of sets as "formulas." Thus one may write

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

or even $P(AB)$ for $P(A \wedge B)$.

One of the most important – but trivial! – theorems is Bayes' theorem.

Theorem 1.7 *Given a probability space* (S, P) *and events* A, B *we have*

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

The proof is immediate from the definitions. Why is such a trivial theorem so important? It is helpful to rewrite it using different letters:

$$P(H|O) = \frac{P(O|H) \cdot P(H)}{P(O)}.$$

where H represents a hypothesis and O represents an observation. What makes this theorem interesting is how it is used in statistical inference;

see any good book on statistical decision theory, for example, the one by Berger [Ber80].

In terms of conditional probability, independence just means that

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

In other words the knowledge that the event B has occurred says nothing about A.

1.3 The Need for Measure Theory

The typical concepts that one learns: independence, expectation value, and conditional probability are fairly clear – at least in their intuitive conception – in the "discrete" case described above. These concepts suffice to analyse much of the work in probabilistic process algebra. In some sense the relevant concepts are essentially those of Boolean algebra. However, in the continuous case, the same concepts require different mathematics. In some sense one can say that one has to move from Boolean algebras to σ-Boolean algebras. Measure theory evolved originally to make sense of integration but – essentially in Kolmogorov's hands – it became the rigorous foundation for probability theory. The need for such extensions to high-school probability theory arose from statistical mechanics and the need to explain physical phenomena like Brownian motion.

For researchers interested in systems like process control systems, telecommunication systems and networks there are very similar phenomena. There is a uncontrolled physical phenomenon, "noise" or "drift," and some controlling software. Understanding how these interact is essential for the design and analysis of such systems. Reasoning under uncertainty is an essential part of Artificial Intelligence and Robotics.

In order to see how measure theory is forced upon us we will consider a classic example – an infinite sequence of coin tosses. This is paradigmatic of an infinitely repeated operation and will be relevant for any analysis of recursive or indefinite iteration in a probabilistic setting. Even though the basic actions are discrete we are led to measure theory by the infinite repetition. Now if we asked naive questions such as "what is the probability of the sequence $(HT)^{\infty}$" we would get 0 as the answer. From this alone we can conclude very little. Right away we observe a striking difference from the finite case. Knowing all the singleton probabilities does not tell us the probabilities associated with other sets. The singleton sets are no longer the "atomic building blocks" from which everything else can be built. We want to be able to say things like "the probability of getting infinitely many heads is 1" which we certainly cannot conclude from simple counting arguments.

What we need is a suitable notion which allows us to define the probabilities in a suitably limiting fashion. We expect that there are certain sets that we can easily associate probabilities with and such that we can define the probabilities associated with other sets by operations. But what are the reasonable operations? It seems compelling that the operations of the discrete theory should survive – these are finite union, finite intersection and complementation. Thus we expect that we will have a family of sets closed under these operations. We further expect that

$$P(A \wedge B) = P(A \cap B)$$

with similar formulas for disjoint union and complementation. We have seen that we cannot expect a summation formula for *arbitrary* unions but, if we want limits to be computable, we can demand that *countable* unions behave like finite unions. In other words we demand that the family of sets we are working with be closed under countable union and complement; intersection is, of course, superfluous. We demand that if we have a pairwise disjoint family of sets A_i, then

$$P(\cup_i A_i) = \sum_i P(A_i) \text{ and } P(A^c) = 1 - P(A).$$

From this, can we compute the probability of having infinitely many heads? The probability of having the first toss be a head followed by an infinite sequence of tails is 0. The probability of exactly one head anywhere is again 0, by considering the countable union. The probability of any fixed finite number of heads is 0, again by taking a countable union and the probability of finitely many heads is again 0. Thus the probability of infinitely many heads is 1. Of course not all answers should be 0 and 1. The probability of a head followed by an *arbitrary* sequence should be 1/2. The sets which look like initial finite sequences followed by an arbitrary sequence are the sets which serve as the basis from which to compute all probabilities.

This raises the natural questions: can we compute probabilities for all the sets this way? It turns out that the answer is no! There are sets for which probability or "measure" cannot be sensibly defined. This never happens when the space of outcomes or states is countable but happens in "almost" any uncountable space.

The key point to take away from this is that we expect to work with *countable* operations – finite ones are not enough and arbitrary ones are impossible.

1.4 The Laws of Large Numbers

One of the most striking early results in probability theory was Borel's law of large numbers. In fact there are two: the **strong** law and the **weak** law . It is worth understanding what they say to see what the difference between discrete and continuous probabilities is. The weak law can be explained (and proved) entirely in terms of discrete probability, but the **strong** law requires the ideas of continuous probability distributions. One can think of this law as corresponding to the choice of a real number from the unit interval or as an infinite sequence of coin tosses. By viewing a sequence of heads and tails as the binary encoding of a real number we can relate the two[2]. We will work with the coin tossing interpretation.

We imagine that we have a fair coin – it has equal probabilities of producing a head or a tail – and we toss it infinitely often. We have seen some of the questions that we can ask about this situation, now we consider subtler questions connected with deviations. Intuitively, we expect that the proportion of heads and tails should be half each, at least in some limiting sense. How can we formalise this? First we can say that the probability that there are n heads in the first $2n$ tosses tends to $1/2$ as n tends to 0. This can be made more formal in the following way. Instead of heads and tails we work with ones and zeros. We write s for a sequence of coin tosses and we write s_i for the value (1 or 0) of the ith toss. Then the number of heads (= 1s) in the first n tosses can be written as $\sum_{i=1}^{n} s_i$. An elementary argument shows that

$$Pr\{s : \sum_{i=1}^{n} s_i = k\} = \binom{n}{k} \frac{1}{2^n}.$$

The weak law expresses the intuition in terms of deviations from the ex-

[2] Modulo some minor niggling details about representing numbers uniquely.

pected probability. For any $\epsilon \geq 0$,

$$\lim_{n \longrightarrow \infty} Pr\{s : |\frac{1}{n}\sum_{i=1}^{n} s_i - \frac{1}{2}| \geq \epsilon\} = 0.$$

This law is expressed entirely in terms of concepts that arise in discrete probability even though it says something nontrivial about a non-discrete situation. The proof can be found in many books; Billingsley [Bil95] or Breiman [Bre68] are good places to look.

We include the proof for the interested reader. The knowledgeable reader can skip this and the beginning reader may defer this to later. For the present we assume that the reader is familiar with the concept of expectation value of a random variable and of the indicator or characteristic function of a set. These are explained later in the text. We proceed as follows. First we prove Chebyshev's inequality for the case of the space of sequences. Let Ω_n be the space of sequences of length n of heads and tails.

Proposition 1.1 ***[Chebyshev]*** *Let X be a function from Ω_n to the reals. Let $\mathsf{E}[X]$ stand for the expectation value of X.*

$$\forall \epsilon > 0. Pr(\{s : X(s) \geq \epsilon\}) \leq \frac{\mathsf{E}[X^2]}{\epsilon^2}.$$

Proof.

$$\begin{array}{ll} & Pr(\{s : X(s) \geq \epsilon\}) \\ = & \mathsf{E}[1_{\{s:X(s)\geq\epsilon\}}] \\ \leq & \mathsf{E}[\frac{X^2}{\epsilon^2}1_{\{s:X(s)\geq\epsilon\}}] \\ \leq & \mathsf{E}[\frac{X^2}{\epsilon^2}] \\ = & \frac{1}{\epsilon^2}\mathsf{E}[X^2] \end{array}$$

□

Proof. (Of the weak law) We define the function $X_j : \Omega_n \longrightarrow \mathbf{R}$ by $X_j(s) = 1$ if the jth element of the sequence s is a head and 0 if it is a tail. We define the function $S_n : \Omega_n \longrightarrow \mathbf{R}$ by $S_n(s) := \sum_{i=1}^{n} X_i(s)$. Thus S_n counts the number of heads in s. Now we can proceed as follows. Note that by applying Chebyshev's inequality with the function $\frac{1}{n}S_n - \frac{1}{2}$ we have immediately

$$Pr(\{s : |\frac{1}{n}S_n - \frac{1}{2}| \geq \epsilon\}) \leq \mathsf{E}[|\tfrac{1}{n}S_n - \tfrac{1}{2}|^2]/\epsilon^2.$$

We can write $\frac{1}{n}S_n - \frac{1}{2}$ as $\frac{1}{n}\sum_{i=1}^{n}(X_i - \frac{1}{2})$. Thus the expectation value that we need is $\frac{1}{n^2}\mathsf{E}[(\sum_{i=1}^{n}(X_i - \frac{1}{2}))^2]$. To calculate the expectation value we first note that if i and j are not equal, then $\mathsf{E}[(X_i - \frac{1}{2})(X_j - \frac{1}{2})] = 0$. Thus when we square the sum, all the cross terms have zero expectation. Thus we are left with $\mathsf{E}[\sum_{i=1}^{n}(X_i - \frac{1}{2})^2]$; each term of this sum has expectation $\frac{1}{4}$ so the sum has expectation $\frac{n}{4}$. Thus

$$Pr(\{s : |\frac{1}{n}S_n - \frac{1}{2}| \geq \epsilon\}) \leq \frac{1}{4n\epsilon^2}.$$

Now when we take the limit $n \longrightarrow \infty$ we get the result. □

Notice that this proof involves purely finite quantities and discrete probability theory.

The so-called *strong* law states something about the probability of sequences that satisfy a condition stated in terms of the entire infinite sequence. We define the set of interest as follows

$$A = \{s : \lim_{n\longrightarrow\infty}[\frac{1}{n}\sum_{i=1}^{n} s_i] = \frac{1}{2}\}.$$

This is the set of sequences with asymptotically equal numbers of heads and tails. Note that the condition of asymptotic equality has to be satisfied by every sequence in the set. When we view these sequences as numbers we get the set of numbers with equal occurrences of the two bits 1 and 0 in their binary representation. Such numbers are called *normal* numbers base 2. One can similarly define normal numbers for any base. A *normal number* is normal with respect to any base. Borel's normal number theorem says that the probability of choosing a non-normal number at random is 0. This is a statement that makes no sense unless we have a notion of *measure* on the entire sets of infinite sequences. We will prove the strong law after we have set up the framework of measure theory.

The strong law implies the weak law but not vice-versa. There are two different notions of convergence at work here. In the weak law we have convergence of the expected values whereas in the strong law we have exact convergence holding with probability 1; what is often called *almost sure* convergence.

1.5 Borel-Cantelli Lemmas

This section is not needed for anything that follows but it helps flesh out the discussion of infinite sequences of coin tosses.

In the last section we considered situations with infinitely many occurrences of an event. In this section we describe two classical lemmas about this type of situation. Suppose we have a situation, say discrete for simplicity, where there are infinitely many events of interest. We write $\{A_i : i \in \mathbf{N}\}$ for these events. The situation that we have in mind is the following. We repeat the experiment infinitely often. In each repetition one (or more) of the A_i may occur or perhaps none of them occur. Now how do we describe the situation "the A_i happened infinitely often" in repeated trials? We are really looking at the countable product space of sequences of trials as we did with the discussion of infinite sequences of heads and tails. The **lim sup** of the sets A_i is given by

$$A = \limsup_{n \longrightarrow \infty} A_n = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n.$$

This is the set theoretical analogue of the "lim sup" which you might remember from undergraduate analysis. It corresponds to the situation we wish to describe, i.e. infinitely often one of the A_i happens. We can now state the first Borel-Cantelli lemma.

Proposition 1.2 ***[Borel-Cantelli I]*** *With the notation above, if* $\sum_{n=1}^{\infty} Pr(A_i) < \infty$ *then* $Pr(A) = 0$.

Proof. Let $B_N = \cup_{n=N}^{\infty} A_n$, then $A = \cap_{N=1}^{\infty} B_N$. Now we have $\forall N, A \subseteq B_N$, thus

$$Pr(A) \leq Pr(B_N) \leq \sum_{n=N}^{\infty} Pr(A_n).$$

Now if we let N go to infinity the sum on the right hand side must go to 0 since it is the tail of the convergent sequence $\sum_{n=1}^{\infty} Pr(A_n)$. Thus $Pr(A) = 0$. □

This makes no assumption about the events being independent. The converse result does, however, require an independence hypothesis.

Proposition 1.3 ***[Borel-Cantelli II]*** *If the A_n are independent then*

$$\sum_{n=1}^{\infty} Pr(A_n) = \infty \textit{ implies } Pr(A) = 1.$$

Proof. Let us write C_n for the complement of A_n, thus $Pr(C_n) = 1 - Pr(A_n)$. Note the obvious inequality $(1 - x) < e^{-x}$ for any x in $(0, 1)$. For any N and $M > N$ we have:

$$\begin{array}{ll} & Pr(\bigcap_{n=N}^{\infty} C_n) \\ \leq & Pr(\bigcap_{n=N}^{M} C_n). \end{array}$$

Now, since we have assumed independence, we have that the last line

$$\begin{array}{ll} = & \prod_{n=N}^{M} Pr(C_n) \\ = & \prod_{n=N}^{M} (1 - Pr(A_n)) \\ \leq & \prod_{n=N}^{M} \exp(-Pr(A_n)) \\ = & \exp(-\sum_{n=N}^{M} Pr(A_n)). \end{array}$$

We used independence in the first equality above. Now if we let M go to ∞ the rhs goes to 0 since the exponent goes to $-\infty$ by hypothesis. Now we have

$$A^c = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} C_n$$

hence

$$Pr(A^c) \leq \sum_{N=1}^{\infty} Pr(\bigcap_{n=N}^{\infty} C_n) = 0.$$

Thus we have $Pr(A) = 1$. □

If we assume independence the result is that the probability of A is always either 1 or 0. Consider our coin-tossing example. If the probability of heads is some fixed number greater than 0 the probability that we will have a sequence of 1729 consecutive heads is 1! Thus even though we expect the numbers to average out "in the long run", in any given sequence of repeated experiments we expect, with high probability, that there are arbitrary long sequences of heads.

Chapter 2

Measure Theory

In this chapter we discuss the axioms for measure theory from an abstract point of view. This will prepare us for integration theory and the major applications of these ideas: to probability theory on continuous state spaces. Intuitively, a measure is a notion of "size" that one wishes to attach to sets. This notion is intended to reflect the geometric notion of size coming from examples like area and volume. Measure turns out to be poorly related to set theoretic conceptions of size. It would be pleasant if we could take all sets to be measurable; unfortunately, as we shall see below, this is not possible, even for such common spaces as **R**. In situations with a countable set of possible states we can indeed take all sets to be measurable and much of the subtleties of measure theory can be dispensed with. However, results and proofs obtained in the discrete case are not very reliable guides to the continuous case.

In this chapter we have given most proofs in greater detail than in most books and we have assumed little prior exposure to advanced analysis. We do assume basic notions in topology and metric spaces.

2.1 Measurable Spaces

Definition 2.1 A **measurable space** (X, Σ) is a set X together with a family Σ of subsets of X, called a **σ-algebra** or **σ-field**, satisfying the following axioms:

(1) $\emptyset \in \Sigma$,
(2) $A \in \Sigma$ implies that $A^c \in \Sigma$, and
(3) if $\{A_i \in \Sigma | i \in I\}$ is a countable family, then $\cup_{i \in I} A_i \in \Sigma$.

If we require only finite additivity rather than countable additivity we get a **field** or **algebra**[1].

Notice that unlike open sets in a topology, measurable sets are closed under complementation and hence under countable intersections as well. This makes a dramatic difference to the properties of measurable functions, compared with continuous functions, as we shall see below. Note also that singletons may or may not belong to a σ-algebra. In most σ-algebras that we are interested in the singletons will be measurable sets.

Here we develop some of the basic properties of σ-algebras.

Proposition 2.1 *The intersection of an arbitrary collection of σ-algebras on a set X is a σ-algebra on X.*

The proof is left as an exercise. The following corollary is very important.

Corollary 2.1 *Given any subset $\mathcal{B}$ of $\mathcal{P}(X)$ there is a least σ-algebra containing $\mathcal{B}$.*

We often refer to the least σ-algebra containing $\mathcal{B}$ as the σ-algebra **generated** by $\mathcal{B}$ and we write $\sigma(\mathcal{B})$ for the σ-algebra generated by $\mathcal{B}$.

Example 2.1 Given a set X the powerset $\mathcal{P}(X)$ is a σ-algebra. The set consisting of just X and $\emptyset$ is another σ-algebra. If X is a countable set and all singletons are measurable the σ-algebra is $\mathcal{P}(X)$. This is the case in most discrete situations.

These are extreme examples of course. A more interesting example and a good source of counter-examples is the following.

Example 2.2 Let X be an uncountable set. The collection of all countable (finite or infinite) sets and cocountable sets (complements of countable sets) forms a σ-algebra on X. This is the σ-algebra generated by the singletons.

The next example is of fundamental importance.

Example 2.3 Given a topological space $(X, \mathcal{T})$ we define $\mathcal{B}(X)$ to be the σ-algebra generated by the open sets (or, equivalently, by the closed sets). Strictly speaking we should write $\mathcal{B}(X)$ since the σ-algebra depends on the topology and not just on the set X, but it is customary to write as we have done since the topology is usually clear from context. The sets in the σ-algebra $\mathcal{B}(X)$ are called **Borel sets**. The most important instance

[1] These words are used differently in abstract algebra.

of this is the collection of Borel sets in $\mathbf{R}$. One often says “Borel sets” to refer to this special case.

The following propositions will be useful when we discuss measurable functions. The proofs are exercises.

Proposition 2.2 *Let $f : X \longrightarrow Y$ be a function and let Σ be a σ-algebra on Y. The sets of the form $\{f^{-1}(A) | A \in \Sigma\}$ form a σ-algebra on X.*

Quite often we want to work with the members of a set which generates a σ-algebra rather than with all sets in the σ-algebra.

Proposition 2.3 *Suppose X, Y are sets and $f : X \longrightarrow Y$ is a function. Suppose that $\mathcal{G}$ is a family of subsets of Y and Σ is the σ-algebra generated by $\mathcal{G}$. Then $\{f^{-1}(A) | A \in \Sigma\}$ is the σ-algebra generated by $\{f^{-1}(A) | A \in \mathcal{G}\}$.*

A categorical interlude

The collection of measurable spaces and measurable functions forms a category which we shall call **Mes**. There is an evident forgetful functor from **Mes** to **Set**. In this category products are constructed as follows. Suppose that (X_1, Σ_1) and (X_2, Σ_2) are measurable spaces. The underlying set is just the cartesian product $X_1 \times X_2$. If $A \in \Sigma_1$ and $B \in \Sigma_2$ are measurable sets the product $A \times B$ is called a rectangle. The σ-algebra on $X_1 \times X_2$ generated by all the rectangles is written $\Sigma_1 \otimes \Sigma_2$. We claim that this is the categorical product in **Mes**.

The situation is illustrated in Fig. 2.1. Given measurable functions f, g from a measurable space Y as shown, the map $\langle f, g\rangle$, constructed as in **Set**, $\langle f, g\rangle(y) = \langle f(y), g(y)\rangle$ is the unique measurable function making the diagram commute. It is in fact the unique function in **Set** making the diagram commute. The only thing to check is that the map is measurable. We leave this as an exercise. **End of categorical interlude.**

The collection of sets in a σ-algebra can be very complicated and one often wants different ways of getting a handle on these sets. Two very powerful theorems that we will use are the *monotone class theorem* and the $\lambda - \pi$ *theorem.* Each of these theorems give alternate ways of getting to a σ-algebra and are useful in showing that various constructions yield σ-algebras. We begin with the monotone class theorem.

The class of σ-algebras can be characterised in terms of monotonicity properties; this will be very useful when we discuss integration. We introduce the following convenient notation. If we have a nested family of

$$\begin{tikzcd}
 & (Y,\Sigma) \arrow[dl] \arrow[dr] \arrow[d, dotted, "\langle f,g\rangle"] & \\
\cdot \arrow[d, "f"] & (X_1 \times X_2, \Sigma_1 \otimes \Sigma_2) \arrow[dl, "\pi_1"'] \arrow[dr, "\pi_2"] & \cdot \arrow[d, "g"] \\
(X_1,\Sigma_1) & & (X_2,\Sigma_2)
\end{tikzcd}$$

Fig. 2.1 Universality diagram for products in **Mes**.

sets

$$A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \ldots$$

with $\cup_n A_n = A$ we write $A_n \uparrow A$. Similarly if

$$A_1 \supseteq A_2 \supseteq \ldots \supseteq A_n \ldots$$

with $\cap_n A_n = A$ we write $A_n \downarrow A$.

Definition 2.2 A collection of sets $\mathcal{M}$ is called a **monotone class** if whenever $A_n \uparrow A$ with all $A_n \in \mathcal{M}$ then $A \in \mathcal{M}$ and also if $A_n \downarrow A$ with all $A_n \in \mathcal{M}$ then $A \in \mathcal{M}$.

Clearly any σ-algebra is a monotone class and, just as for σ-algebras, the intersection of monotone classes is a monotone class. Thus we can talk about the monotone class generated by a collection of sets just as we did for σ-algebras. Recall that a field is like a σ-algebra except that we only require finite additivity rather than countable additivity.

Proposition 2.4 *Any σ-algebra is a monotone class and if a monotone class is also a field then it is a σ-algebra.*

Proof. Suppose that Σ is a σ-algebra on X and $A_i \uparrow A$ with $A_i \in \Sigma$. Then $A = \bigcup_i A_i$, so $A \in \Sigma$. If $A_i \downarrow A$ with $A_i \in \Sigma$, then $A_i^c \uparrow A^c$ so A^c is in Σ and hence A is in Σ. Thus Σ is a monotone class.

Suppose that $\mathcal{M}$ is a monotone class and a field. Since $\mathcal{M}$ is a field $\emptyset$ and X are in $\mathcal{M}$ and $\mathcal{M}$ is closed under complementation. It remains to show that $\mathcal{M}$ is closed under countable unions. Let $\{A_i | i \in \mathbf{N}\}$ be a countable family of sets in $\mathcal{M}$. Define $B_i = \bigcup_{j=1}^{i} A_i$; since $\mathcal{M}$ is a field all the B_i are in $\mathcal{M}$. Clearly $B_i \uparrow (\bigcup_j A_j)$ hence $\bigcup_j A_j \in \mathcal{M}$. □

Theorem 2.1 *If $\mathcal{F}$ is a field of subsets of X then the monotone class, $m(\mathcal{F})$, generated by $\mathcal{F}$ is the same as $\sigma(\mathcal{F})$.*

Proof. It suffices to show that $m(\mathcal{F})$ is a field; the above proposition would then imply that $m(\mathcal{F})$ is a σ-algebra. Since $\sigma(\mathcal{F})$ is the smallest σ-algebra containing $\mathcal{F}$ we would then have $\sigma(\mathcal{F}) \subseteq m(\mathcal{F})$; furthermore since any σ-algebra is a monotone class and $m(\mathcal{F})$ is the smallest monotone class containing $\mathcal{F}$ we have $m(\mathcal{F}) \subseteq \sigma(\mathcal{F})$.

First we show that $m(\mathcal{F})$ is closed under complementation. We define $\mathcal{N} := \{E \in m(\mathcal{F}) | E^c \in m(\mathcal{F})\}$. Clearly $\mathcal{F} \subseteq \mathcal{N}$. Now suppose that $A_i \uparrow A$ with all A_i in $\mathcal{N}$. Then A_i^c are all in $m(\mathcal{F})$ with $A_i^c \downarrow A^c$, thus – since $m(\mathcal{F})$ is a monotone class – we have $A^c \in m(\mathcal{F})$ or $A \in \mathcal{N}$. Similarly we can show that if $A_i \downarrow A$, where all the A_i are in $\mathcal{N}$, then $A \in \mathcal{N}$. Thus $\mathcal{N}$ is a monotone class containing $\mathcal{F}$. Since $m(\mathcal{F})$ is the smallest monotone class containing $\mathcal{F}$ we have $m(\mathcal{F}) \subseteq \mathcal{N}$; by definition $\mathcal{N} \subseteq m(\mathcal{F})$ so $\mathcal{N} = m(\mathcal{F})$ which means that $m(\mathcal{F})$ is closed under complementation.

For each $A \subset X$ define $M_A := \{E | E \cap A \in m(\mathcal{F})\}$. Suppose that $B_i \uparrow B$, with all the B_i in M_A, then $(B_i \cap A) \uparrow (B \cap A)$. Since $(B_i \cap A) \in m(\mathcal{F})$ we have $(B \cap A) \in m(\mathcal{F})$ so $B \in M_A$. Similarly the other direction can be established and it follows that M_A is a monotone class.

Now clearly $A \in M_B$ if and only if $B \in M_A$. If $A \in \mathcal{F}$, then M_A is a monotone class containing $\mathcal{F}$, hence $m(\mathcal{F}) \subset M_A$. Let $E \in m(\mathcal{F})$ and $A \in \mathcal{F}$, then $E \in M_A$, hence $A \in M_E$, so M_E is a monotone class containing $\mathcal{F}$ and $m(\mathcal{F}) \subseteq M_E$. Thus given $D, E \in m(\mathcal{F})$ we have $D \in M_E$ which means $D \cap E \in m(\mathcal{F})$. Thus $m(\mathcal{F})$ is closed under finite intersection and we have completed the proof that $m(\mathcal{F})$ is a field. □

The sets in a σ-algebra may be complicated. It is often advantageous to work with a generating collection. For example, it is hard to describe general Borel sets but it is easy to describe open sets. A particularly simple kind of family of sets is called a π-system.

Definition 2.3 A π-system is a family of sets closed under finite intersections.

The open intervals of $\mathbf{R}$ form a π-system and they generate the Borel sets. A structure similar to a σ-algebra is called a λ-system.

Definition 2.4 A λ-system over a set X is a family of sets (i) containing X, (ii) closed under complementation and (iii) closed under countable unions of pairwise disjoint sets.

Like σ-algebras the intersection of any collection of λ-systems is a λ-system so one can talk sensibly about the λ-system generated by a family of sets.

The following results show the relation between π-systems and λ-systems.

Proposition 2.5 *If $\mathcal{P}$ is both a π-system and a λ-system then it is a σ-algebra.*

Proof. We have to show that $\mathcal{P}$ is closed under arbitrary countable unions. Let A_n be a countable family of sets in $\mathcal{P}$. We define $B_n = A_n \cap (\bigcap_{i=1}^{n-1} A_i^c)$. Since $\mathcal{P}$ is a λ-system the $(A_i)^c$ are in $\mathcal{P}$ and since $\mathcal{P}$ is a π-system the B_n are in $\mathcal{P}$. Suppose that $x \in \cup_i A_i$; let j be the smallest index such that $x \in A_j$, then $x \in B_j$ but not in any other B_j so they are all pairwise disjoint. Since $\mathcal{P}$ is a λ-system we have that $\cup_i B_i \in \mathcal{P}$. Clearly $\cup_i A_i = \cup_i B_i$ so $\mathcal{P}$ is closed under arbitrary countable unions. □

The main theorem is the following, known as Dynkin's $\lambda - \pi$ theorem.

Theorem 2.2 *If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system then $\mathcal{P} \subset \mathcal{L}$ implies that $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

Proof. Let $\lambda(\mathcal{P})$ be the λ-system generated by $\mathcal{P}$. The strategy is to prove that $\lambda(\mathcal{P})$ is a π-system and hence a σ-algebra. From this it follows that

$$\mathcal{P} \subset \sigma(\mathcal{P}) \subseteq \lambda(\mathcal{P}) \subseteq \mathcal{L}.$$

Let $L_A := \{B \subseteq X | A \cap B \in \lambda(\mathcal{P})\}$. Now we claim that if $A \in \lambda(\mathcal{P})$ then L_A is a λ-system. To see this we verify the properties of an λ-system directly.

$A \cap X = A \in \lambda(\mathcal{P})$ so X is in L_A. If $A \in \lambda(\mathcal{P})$ then clearly $A \in L_A$.

Let $B \in L_A$, i.e. $A \cap B \in \lambda(\mathcal{P})$. We need to show that B^c is also in L_A, i.e. $A \cap B^c \in \lambda(\mathcal{P})$. We have assumed that $A \in \lambda(\mathcal{P})$. Now A^c is disjoint from $A \cap B$ so the union $A^c \cup (A \cap B)$ is in $\lambda(\mathcal{P})$. The complement of this, which is $A \cap B^c$, is also in $\lambda(\mathcal{P})$. Thus B^c is in L_A.

Let B_n be a pairwise disjoint family in L_A then $A \cap B_n$ is a pairwise disjoint family in $\lambda(\mathcal{P})$ so $\cup_n (A \cap B_n) = A \cap (\cup_n B_n)$ is in $\lambda(\mathcal{P})$, hence $\cup_n B_n$ is in L_A. Thus L_A is a λ-system.

Suppose that $A, B \in \mathcal{P}$, since $\mathcal{P}$ is a π-system $A \cap B \in \mathcal{P} \subset \lambda(\mathcal{P})$ so $B \in L_A$. Thus $\mathcal{P} \subseteq L_A$ and hence $\lambda(\mathcal{P}) \subseteq L_A$. If $A \in \mathcal{P}$ and $B \in \lambda(\mathcal{P})$ then $B \in L_A$ so $A \cap B \in \lambda(\mathcal{P})$ which means $A \in L_B$. Thus, for $B \in \lambda(\mathcal{P})$, $\mathcal{P} \subseteq L_B$ i.e. $\lambda(\mathcal{P}) \subseteq L_B$. Now suppose $A, B \in \lambda(\mathcal{P})$, we have $\lambda(\mathcal{P}) \subseteq L_B$, so $A \in L_B$, hence $A \cap B \in \lambda(\mathcal{P})$. Thus, indeed $\lambda(\mathcal{P})$ is a π-system, hence a σ-algebra. □

When we introduce measures π-systems will play a crucial simplifying role.

2.2 Measurable Functions

Propositions 2.2 and 2.3 show that σ-algebras behave well under inverse images. Accordingly it is natural to define measurable functions in terms of inverse images.

Definition 2.5 A function f from a σ-algebra (X, Σ_X) to a σ-algebra (Y, Σ_Y) is said to be **measurable** if $f^{-1}(B) \in \Sigma_X$ whenever $B \in \Sigma_Y$.

This parallels the definition of continuous function in topology. Traditionally the phrase "measurable function" is used for a real-valued function but we will use it more generally. Our measurable functions have been sometimes called "measurable transformations". If we consider topological spaces with their Borel σ-algebra, then any continuous function is clearly measurable. However many discontinuous functions are also measurable.

2.3 Metric Spaces and Properties of Measurable Functions

We now consider some special spaces and measurable functions to them. The most important measurable space is **R** with the Borel algebra generated by the usual topology. Unless we say otherwise we shall always mean "reals equipped with this topology and σ-algebra" when we refer to the reals. The following seemingly easy proposition gives some closure properties of real-valued measurable functions; it will turn out to conceal a subtlety that bears further scrutiny.

Proposition 2.6 *The absolute value of a measurable function is measurable. The sum and product of two measurable functions are measurable. The multiplicative inverse of an everywhere nonzero measurable function is measurable.*

Proof. The absolute value function is continuous and hence measurable, hence the first statement is proved. For the next statement the same argument should work since $+, * : \mathbf{R}^2 \longrightarrow \mathbf{R}$ are continuous. However, what this proves is that $+, *$ are measurable with respect to the product Borel algebra $\mathcal{B} \otimes \mathcal{B}$ on $\mathbf{R}^2$, which is not a priori the same as the Borel algebra generated by the product topology. In fact it is, but the proof is not trivial and is the subject of the next proposition. To verify the last statement let $\mathbf{R}'$ be the reals with zero removed; let $\eta : \mathbf{R}' \longrightarrow \mathbf{R}$ be given by $\eta(x) = \frac{1}{x}$. Clearly η is continuous. Given f a nonzero measurable function we define $g(x) = \frac{1}{f(x)}$. According to proposition 2.3 we need to verify that $g^{-1}(O)$ is measurable only for an open set O in $\mathbf{R}$. Now we have $g^{-1}(O) = g^{-1}(O \cap \mathbf{R}') = f^{-1}(\eta^{-1}(O \cap \mathbf{R}'))$; this is a measurable set since η is continuous and f is measurable. □

In order to fill the lacuna in the above proof we need a digression into metric spaces. Metric spaces play a central role in probability theory though one would not have guessed that from the initial definitions.

Definition 2.6 A metric space with a countable, dense subset is called **separable**.

Typical examples are $\mathbf{R}^n$. What separability does is give one access to a *countable* family of generating open sets

Proposition 2.7 *In a separable metric space, say (X, d), there is a countable family of open sets H_i such that every open subset may be written as the union of the H_i that it contains.*

Proof. Let $\{x_i\}$ be a dense subset of (X, d). Define $H_{i,m} = \{x | d(x, x_i) < 1/m\}$, where m is an integer. Let O be an open subset of X and let z be a point in O. Now there exists an integer n such that the open ball centred at z and with radius $1/n$ is contained in O. Since the set $\{x_i\}$ is dense there is a j such that $d(z, x_j) < 1/(2n)$. Then $H_{j,2n}$ is contained in O and includes the point z. □

Now we can relate the two Borel structures as required.

Proposition 2.8 *Let (X_1, d_1) and (X_2, d_2) be two separable, metric spaces. Let Y be the product equipped with the product topology. Then the Borel algebra of Y, written $\mathcal{B}_Y$, is the same as the product σ-algebra $\mathcal{B}_1 \otimes \mathcal{B}_2$.*

Proof. It is easy to see that Y is also separable; in fact we can construct the countable family of open sets by taking products of open balls. Thus we have a countable family of open rectangles which generate the topology of Y and hence the Borel algebra $\mathcal{B}_Y$. These are all in the σ-algebra $\mathcal{B}_1 \otimes \mathcal{B}_2$. Now any open set is the *countable* union of these special rectangular open sets and thus any open set in the product topology of Y is in $\mathcal{B}_1 \otimes \mathcal{B}_2$. Thus we have that $\mathcal{B}_Y \subseteq \mathcal{B}_1 \otimes \mathcal{B}_2$. Now consider the identity function on Y viewed as a function between measurable spaces, i.e. $i_Y : (Y, \mathcal{B}_Y) \longrightarrow (Y, \mathcal{B}_1 \otimes \mathcal{B}_2)$. This function is the one induced by universality of the product given the projections onto the components. Clearly the projections are continuous (by definition of the product topology) and hence measurable, thus the map i_Y is measurable. This means that $\mathcal{B}_1 \otimes \mathcal{B}_2 \subseteq \mathcal{B}_Y$. □

We now discuss a theorem which shows the striking contrast between measure theory and topology. We take (X, Σ) to be a measurable space, (Y, d) to be a metric space with the induced Borel algebra $\mathcal{B}_Y$.

Definition 2.7 Given a family of functions $\{f_n : X \longrightarrow Y | n \in \mathbb{N}\}$ we say that the family **converges pointwise** to f if $\forall x \in X. \lim_{n\to\infty} f_n(x) = f(x)$.

Theorem 2.3 *If a family of measurable functions* $\{f_n : X \longrightarrow Y | n \in \mathbb{N}\}$ *converges pointwise to* f*, then* f *is also measurable.*

Proof. We need to show that for an open set O of Y that $f^{-1}(O)$ is measurable. We will describe this set in terms of the f_n and use measurability of the f_n to establish measurability of f.

Given a subset O of Y and a point $y \in Y$ we define a distance function, also written d, by $d(y, O) = \inf_{u \in O} d(u, y)$. Let O be an open set in Y and let k be a positive integer, we define

$$O_k = \{u \in O | d(u, O^c) > 1/k\}.$$

Now it is not hard to see that the O_k are open sets[2], so we have an increasing sequence of open sets contained in O with $O = \cup_{k \in \mathbb{N}} O_k$. In fact we have $\overline{O_k} \subseteq O_{k+1}$.

Now fix a point x of X. Suppose that x is in $f^{-1}(O)$, then x has to be in $f^{-1}(O_k)$ for some k. We have $\lim_{n\to\infty} d(f_n(x), f(x)) = 0$ by hypothesis. Thus, for q large enough, say $\forall q \geq m$, we have $f_q(x) \in O_k$. Let $H^k_m = \cap_{q \geq m} f_q^{-1}(O_k)$. This set is measurable since all the f_n are and the countable intersection of measurable sets is measurable. Clearly we have $f^{-1}(O_k) \subseteq \cup_m H^k_m$. We define $G^k = \cup_m H^k_m$, which is also a measurable set being the countable union of measurable sets. Taking the union over all k gives $f^{-1}(O) = \cup_k f^{-1}(O_k) \subseteq \cup_k G^k \overset{\text{def}}{=} W$. Clearly W is measurable. Now we show that W is in fact $f^{-1}(O)$ to complete the argument. Let $u \in W$,

[2] It is easy to verify that the complements are closed sets.

i.e. there is an r such that $u \in G^r$, i.e. there is a j such that $u \in H^r_j$. Thus for all $q > j$ we have $u \in f_q^{-1}(O_r)$. Thus we have

$$f(u) = \lim_{q \to \infty} f_q(u) \in \overline{O_r} \subseteq O_{r+1} \subseteq O.$$

Thus $u \in f^{-1}(O)$ and $W = f^{-1}(O)$, thus f is measurable. □

This is a very typical argument and is worth ploughing through for that reason alone. The discussion can be turned around to show that every measurable function is the pointwise limit of very special functions. This fact lies at the heart of integration theory.

Definition 2.8 A measurable function is called **simple** if its range is finite.

For real valued functions we have the following very important result.

Theorem 2.4 *Given a nonnegative measurable function $f : X \longrightarrow \mathbf{R}$ there is a family of simple functions s_i such that $\forall i . s_i \leq s_{i+1} \leq f$ and the s_i converge pointwise to f.*

Proof. The strategy is to define step functions but on possibly complicated domains obtained by inverse images. For $n \in \mathbb{N}$ and $1 \leq i \leq n2^n$ define $E_{n,i} \stackrel{\text{def}}{=} f^{-1}([\frac{i-1}{2^n}, \frac{i}{2^n}])$ and $F_n \stackrel{\text{def}}{=} f^{-1}([n, \infty])$. Since f is measurable the sets E and F are also measurable. We set

$$s_n = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} \chi_{E_{n,i}} + n \chi_{F_n}.$$

It is easy to see that the s_n are increasing. Suppose that $f(x) = r$ for some x and suppose that $r \in [\frac{i-1}{2^n}, \frac{i}{2^n}]$. Then $x \in E_{n,i}$ and $s_n(x) = \frac{i-1}{2^n} \leq r$. Note that s_n stays within $\frac{1}{2^n}$ of $f(x)$ as n increases so s_n converges to f. If $f(x) = \infty$ for some x, then $s_n(x) = n$ which will diverge as n goes to infinity. □

2.4 Measurable Spaces of Sequences

A very important construction that appears repeatedly in computer science is the construction of a σ-algebra on the space of all finite and infinite sequences over some alphabet. Such spaces of sequences arise as traces of a process and hence it is a very important example. Such sets of sequences are typically turned into topological spaces or metric spaces and hence one can look at the generated σ-algebra. We will, however, construct the σ-algebra directly: this is called the "cone" construction.

First, we fix an alphabet $\mathcal{A}$, which we take to be countable. We write $\mathcal{A}^{\infty}$ for the set of finite and infinite sequences over $\mathcal{A}$. This set comes with some natural structures. First of all it is a monoid with respect to the operation of concatenation of sequences with the empty sequence ϵ as the identity element. We will indicate concatenation by juxtaposition. There is an order structure: $\alpha \sqsubseteq \beta$ is defined to mean that α is a prefix of β.

Now we define a *cone* to be a set of the form

$$(\alpha)\uparrow := \{\beta : \alpha \sqsubseteq \beta\},$$

where α is a finite sequence, i.e. an element of $\mathcal{A}^{*}$. These cones are the base for a topology – the Scott topology – much used in domain theory. The σ-algebra generated by the cones is the one of interest.

What are examples of measurable sets? "All sequences with the symbol a at the seventeenth position" is one example. Why is this measurable? Note that there are countably many finite sequences with length 17 and with an a in the seventeenth position. Take the union of all the cones generated by all these sequences: this is the set defined in the last sentence and, being the countable union of measurable sets, is clearly measurable. Similarly we can fix finitely many positions along the sequence and specify the element appearing there and let all other positions have any element; this too yields a measurable set.

2.5 Measures

Measurable spaces or σ-algebras are merely the arenas in which measure theory happens. The key notion of "measure" will now be introduced. Roughly speaking, a measure is an assignment of size to the sets in σ-algebra. This size is typically a real number but it could be a real number between 1 and 0, a probability measure, or an extended nonnegative real number, i.e. one from $[0, \infty]$, or even a complex number. These theories are all slightly different and play different roles in mathematics. For us the most important case will be probability measure but it is worth seeing what happens when ∞ is admitted as a possible value; this is of particular importance in integration theory.

Definition 2.9 A **measure** (**probability measure**) μ on a measurable space (X, Σ) is a function from Σ (a set function) to $[0, \infty]$ ($[0, 1]$), such that if $\{A_i | i \in I\}$ is a countable family of pairwise disjoint sets, then

$$\mu(\bigcup_{i\in I} A_i) = \sum_{i\in I} \mu(A_i).$$

In particular if I is empty we have

$$\mu(\emptyset) = 0.$$

A set equipped with a σ-algebra and a measure defined on it is called a **measure space**.

This property is called *countable additivity* or *σ-additivity.* It is possible to develop a theory with just finite additivity but many basic results are counterintuitive.

In the rest of this section we will always consider a set X equipped with a σ-algebra Σ and a measure μ. We shall always mean "measurable set" when we just say "set". We use letters like A, B to stand for measurable sets.

Proposition 2.9 ***[Monotonicity and Continuity]***

(1) If $A \subseteq B$ then $\mu(A) \leq \mu(B)$.
(2) If $A_1 \subseteq A_2 \subseteq \ldots A_n \subseteq \ldots$ and $\cup_i A_i = A$ then $\lim_{i\rightarrow\infty} \mu(A_i) = \mu(A)$.
(3) If $A_1 \supseteq A_2 \supseteq \ldots A_n \supseteq \ldots$ and $\cap_i A_i = A$ then $\lim_{i\rightarrow\infty} \mu(A_i) = \mu(A)$, if $\mu(A_1)$ is finite.

Proof. For (1) we note that $B = A \cup (B - A)$. Thus $\mu(B) = \mu(A) + \mu(B - A) \geq \mu(A)$.

For (2) we define a family of pairwise disjoint sets B_i inductively as follows. We set $B_1 = A_1$ and $B_{i+1} = A_{i+1} - A_i$. Since we have $\cup_i B_i = A$ and the B_i are pairwise disjoint, we have, by countable additivity, that $\mu(A) = \sum_i B_i$. But we also have $A_n = \bigcup_{i=1}^n B_i$, thus $\mu(A_n) = \sum_{i=1}^n \mu(B_i)$. Thus

$$\lim_{n\longrightarrow\infty} \mu(A_n) = \lim_{n\longrightarrow\infty} \sum_{i=1}^{n} \mu(B_i) = \sum_{i=1}^{\infty} \mu(B_i) = \mu(A).$$

A similar argument holds for the third clause though the caveat about $\mu(A_1)$ being finite is essential. □

The following corollary is immediate.

Corollary 2.2 ***[Convexity]*** *For any countable family of sets B_i we have $\mu(\cup_i B_i) \leq \sum_i \mu(B_i)$.*

The first example looks natural but is pathological and is an important source of counterexamples.

Example 2.4 For X an infinite set we define a measure on the powerset of X by setting $\mu(A)$ equal to the number of elements of A if A is finite and ∞ otherwise. This measure is called *counting measure.* Many small variations are possible, such as weighting the points of X differently.

The next example appears artificial but is of central importance.

Example 2.5 Fix a set X and a point x of X. We define a measure, in fact a probability measure, on the σ-algebra of all subsets of X as follows. We use the slightly peculiar notation $\delta(x, A)$ to emphasise that x is a parameter in the definition.

$$\delta(x, A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

This measure is called the *Dirac delta measure*. Note that we can fix the set A and view this as the definition of a (measurable) function on X. What we get is the characteristic function of the set A, χ_A.

In order to cope with ∞ as a legitimate value in measure theory we need to adopt some arithmetic conventions for dealing with it. We decree that $0 \cdot \infty = 0$, otherwise we adopt the reasonable rules, $a \cdot \infty = \infty \cdot a = \infty$ if $a \neq 0$ and $a + \infty = \infty + a = \infty$. This choice gives commutativity and associativity of addition and multiplication as well as unrestricted distributivity. Furthermore if we have two sequences $a_n \longrightarrow a$ and $b_n \longrightarrow b$ we will have $a_n \cdot b_n \longrightarrow a \cdot b$ whatever a or b might be.

A measure space (X, Σ, μ) is said to be **finite** if $\mu(X) < \infty$ and **σ-finite** if X can be written as a countable union of sets of finite measure. Finite measure spaces arise in probability theory but many results actually hold for the σ-finite case and in these notes we will not always assume finiteness.

Proving the Strong Law

We now have the technical tools to prove the strong law . First of all recall that the sequence space is Ω. The σ-algebra on Ω is defined as follows. We consider all sets of form

$$A_s := \{s \cdot s' : s \text{ a finite sequence, } s' \in \Omega\}$$

where $\cdot$ stands for concatenation of sequences. We consider the σ-algebra generated by these sets. We assign a probability measure Pr as follows.

For any set of the form A_s we associate the probability $2^{-\text{length}(s)}$.

Recall that we defined the set $A \subseteq \Omega$ by

$$A := \{s : \lim_{n \longrightarrow \infty} [\frac{1}{n} \sum_{i=1}^{n} s_i] = \frac{1}{2}\}.$$

The strong law says that the probability of landing in the set A is 1 if we choose an infinite sequence randomly. It will be more convenient to prove that $E := A^c$ has probability 0.

Proof. (Of the strong law) We define the measurable functions $X_j : \Omega \longrightarrow \mathbf{R}$ by

$$X_j(s) = \begin{cases} 1 & \text{if } s_j = H \\ 0 & \text{if } s_j = T. \end{cases}$$

We define the function $S_n(s) := \sum_{j=1}^{n} X_j(s)$. Then we can define A as the set $\{s : \lim_{n \longrightarrow \infty} \frac{S_n(s)}{n} = \frac{1}{2}\}$.

We claim that $\lim_{n \longrightarrow \infty} \frac{S_n(s)}{n} = \frac{1}{2}$ if and only if $\lim_{m \longrightarrow \infty} \frac{S_{m^2}(s)}{m^2} = \frac{1}{2}$. To see this, for any n we choose m by $m^2 \leq n \leq (m+1)^2$, or equivalently, $0 \leq n - m^2 \leq 2m$. If we choose m this way we have

$$|\frac{S_n}{n} - \frac{S_{m^2}}{m^2}| = |\frac{S_n}{m^2} - \frac{S_{m^2}}{m^2} + (\frac{1}{n} - \frac{1}{m^2})S_n|$$

$$\leq \frac{|n - m^2|}{m^2} + n|\frac{1}{n} - \frac{1}{m^2}| \leq \frac{2}{m} + \frac{2}{m} = \frac{4}{m}.$$

Thus as $n, m \longrightarrow \infty$ the difference between $\frac{S_n}{n}$ and $\frac{S_{m^2}}{m^2}$ goes to 0. Hence we can now work with $\frac{S_{m^2}}{m^2}$ in reasoning about E.

The strategy of the proof is to consider sets of sequences, E_ϵ, where the asymptotic behaviour is bounded away from 0 by ϵ. The set E is a countable union of such sets. We will show that each of the E_ϵ has probability 0, thus E will also have probability 0. To show that each E_ϵ has probability 0 we will use the limiting properties of the probability.

We define $E_\epsilon := \{s : \overline{\lim} |\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon\}$. We define

$$E_a^b := \bigcup_{m=a}^{b} \{s : |\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon\}.$$

The set E_a^b represents the sequences where the inequality $|\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon$ is satisfied at least once for m between a and b. Finally we define

$$E_a = \bigcup_{b=a}^{\infty} E_a^b.$$

By the basic convexity property we have

$$Pr(E_a^b) \leq \sum_{m=a}^{b} Pr(\{s : |\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon\}).$$

Now using Chebyshev's inequality just as we did in the proof of the weak law we get

$$Pr(E_a^b) \leq \frac{1}{4\epsilon^2} \sum_{m=a}^{b} \frac{1}{m^2}.$$

As we let b go to infinity and use the fact that for fixed a the family E_a^b is a nested increasing family of sets, so that by continuity we get

$$Pr(E_a) = \lim_{b \longrightarrow \infty} Pr(E_a^b) \leq \frac{1}{4\epsilon^2} \sum_{m=a}^{\infty} \frac{1}{m^2}.$$

What does it mean for a sequence to be in a set of the form E_a? It means that at some stage *after* a the inequality $|\frac{S_{m^2}}{m^2} - \frac{1}{2}| > \epsilon$ is satisfied. Now, by the definition of $\overline{\lim}$ to be in E_ϵ this inequality must be satisfied infinitely often, i.e. we have $E_\epsilon = \cap_a E_a$. As a increases we get a nested decreasing family of sets so by continuity again we have $Pr(E_\epsilon) = \lim_{a \longrightarrow \infty} Pr(E_a)$. Thus

$$Pr(E_\epsilon) \leq \lim_{a \longrightarrow \infty} [\frac{1}{4\epsilon^2} \sum_{m=a}^{\infty} \frac{1}{m^2}] = 0.$$

The last equality follows from the fact that we are looking at the tails of a convergent sequence. Thus we have established that for any $\epsilon > 0$, $Pr(E_\epsilon) = 0$.

Now consider the set E. This set can be written as the countable union

$$E = \bigcup_{k=1}^{\infty} E_{\frac{1}{k}}$$

where k is an integer. Thus since each $E_{\frac{1}{k}}$ has probability 0 we get that E has probability 0, or A has probability 1. □

A set of measure 0 is sometimes called a **negligible** set. They play a very important role in the theory and one often hears phrases like "almost everywhere" or "almost surely". What they usually mean is that something or other is true except on a negligible set. To be sure the notion of negligibility depends on the measure so one should be careful in interpreting such phrases. One very annoying feature of negligible sets is that they may contain nonmeasurable subsets, whereas we would certainly like to say that all subsets of a negligible set are negligible as well. This would be true if we could be certain that all the subsets of a negligible set were measurable. Fortunately this can be fixed by "completing the measure". One cannot just throw in all subsets of the negligible sets but what turns out to work is to throw in all sets that are sandwiched between two sets of the same measure. We proceed to make all this precise.

Definition 2.10 A measure space (X, Σ, μ) is said to be **complete** if every subset of a negligible set is in Σ.

Theorem 2.5 *Given (X, Σ, μ), there is a σ-algebra, $\Sigma' \supseteq \Sigma$ and a measure μ' on Σ' such that (X, Σ', μ') is complete and $\mu(A) = \mu'(A)$ for any A in Σ.*

Proof. We define the extended σ-algebra as follows

$$\Sigma' = \{Z \in \mathcal{P}(X) | \exists A, B \in \Sigma . A \subseteq Z \subseteq B \wedge \mu(B - A) = 0\}.$$

Clearly $\Sigma' \supseteq \Sigma$, but we need to verify that it is indeed a σ-algebra. Let Z be in Σ'. Then we have A, B as above. Since $A \subseteq Z \subseteq B$ we have $B^c \subseteq Z^c \subseteq A^c$ and, of course, A^c and B^c are both in Σ. Also $\mu(A^c - B^c) = \mu(B - A) = 0$ so Z^c is in Σ'. Now suppose that Z_n is a countable family of sets in Σ'. We have two countable families of sets A_n and B_n both drawn from Σ with

$$\forall n . A_n \subseteq Z_n \subseteq B_n, \mu(B_n - A_n) = 0.$$

We set $Z = \cup_n Z_n$, $A = \cup_n A_n$ and $B = \cup_n B_n$. Clearly A and B are in Σ and $A \subseteq Z \subseteq B$. Now $(B - A) \subseteq \cup_n [B_n - A_n]$ hence by monotonicity of measure and convexity we have

$$\mu(B - A) \leq \mu(\cup_n [B_n - A_n]) \leq \sum_n \mu(B_n - A_n) = 0.$$

Thus Z is in Σ' and we have verified that Σ' is a σ-algebra.

We define μ' as follows. First, note that $A \subseteq B$ and $\mu(B - A) = 0$ implies that $\mu(A) = \mu(B)$. Given $A \subseteq Z \subseteq B$ we define $\mu'(Z) = \mu(A) = \mu(B)$.

We need to verify that this is independent of the chosen A and B. Suppose that we have another pair $A' \subseteq Z \subseteq B'$ with $\mu(B' - A') = 0$. Now since $B' \supseteq Z \supseteq A$ and $B \supseteq Z \supseteq A'$, we have $\mu(B') \geq \mu(A) = \mu(B)$ and $\mu(B) \geq \mu(A') = \mu(B')$. Thus we have $\mu(A) = \mu(A') = \mu(B) = \mu(B')$ and μ' is well defined. Finally we need to verify countable additivity of μ'. Suppose that we have a pairwise disjoint family of sets Z_n from Σ'. Then the associated sets A_n are also pairwise disjoint, thus we have

$$\mu'(\cup_n Z_n) = \mu(\cup_n A_n) = \sum_n \mu(A_n) = \sum_n \mu'(Z_n).$$

To verify that μ' is complete we proceed as follows. Suppose that $Z \in \Sigma'$, $Z' \subseteq Z$ and that we have $\mu'(Z) = 0$. Then we have sets $A, B \in \Sigma$ with $A \subseteq Z \subseteq B$ and with $\mu(B) = 0$. Now we have $\emptyset \subseteq Z' \subseteq B$, thus Z' is in Σ' and has μ'-measure 0. □

Remark 2.1 *The word "completion" is not a very good choice. It suggests that the completion is unique*[3] *but it is not in fact. It is possible to give simple (finite) examples where the completion process can be further extended. What is true is that the completion process described in the proof above, if carried out a second time, would yield no new sets.*

Given a measure space (X, Σ, μ) one often implicitly talks about the completion with respect to μ rather than the given σ-algebra. It is conventional when talking about the reals to use the phrase "Borel field" or "Borel sets" to refer to the σ-algebra generated by the open intervals, and the phrase "Lebesgue measurable sets" to talk about the sets that arise from the completion process above with respect to the standard Lebesgue measure, defined below.

2.6 Lebesgue Measure

We now proceed to construct the most important single example of a measure, the Lebesgue measure on the real line. We want a notion of measure on the real line to assign the familiar notion of length to intervals. This simple requirement immediately brings us face-to-face with the technical subtleties that lie at the foundations of set theory.

Theorem 2.6 ***[Vitali]*** *Assuming the axiom of choice, there is no measure on the real numbers which is translation-invariant, assigns a measure*

[3] Indeed some reputable books say that it is.

to all sets and which assigns a nonzero value to the interval $[0,1]$.

We defer the proof to the end of the section in order not to interrupt the flow of ideas.

Let us start naively. We would like a measure which agrees with the usual notion of length of an interval. Recall that $[0,\infty]$ is a complete lattice so we can take lubs and glbs as we please. Let us call our putative measure m and we begin by trying to define it on all subsets of $\mathbf{R}$. We begin as follows

$$m((a,b)) = m((a,b]) = m([a,b)) = m([a,b]) = b-a$$

where a and b are real numbers. If we have a pairwise disjoint collection of intervals $\{I_k | k \in \mathcal{K}\}$ we define

$$m(\bigcup_{k\in\mathcal{K}} I_k) = \sum_{k\in\mathcal{K}} m(I_k)$$

as suggested by the countable additivity requirement. It is not hard to check that m does not depend on how we partition a set into disjoint intervals. Now suppose that we have an arbitrary subset A of $\mathbf{R}$. We define a putative measure μ^* as follows. Let S be a family of intervals that covers A. We set $\mu^*(A)$ to be the infimum over all families S that cover A, of $m(S)$. This is indeed defined on all subsets of $\mathbf{R}$. Unfortunately it is not a measure because it is not countably additive. Nevertheless it is an interesting set function which will be the basis of our construction of the Lebesgue measure.

As a prelude let us calculate μ^* for some sets. Clearly $\mu^*([a,b]) = b-a$ and $\mu^*(\mathbf{R}) = \infty$. We claim that $\mu^*(\mathbb{Q}) = 0$ where $\mathbb{Q}$ is the set of rational numbers. Let ϵ be any positive number, we will cover the rationals with a family of intervals of total length ϵ. Let $r_1, r_2, \ldots$ be any enumeration of the rationals. Consider the family of intervals

$$\{[r_1 - \frac{\epsilon}{4}, r_1 + \frac{\epsilon}{4}], [r_2 - \frac{\epsilon}{8}, r_2 + \frac{\epsilon}{8}], \ldots, [r_j - \frac{\epsilon}{2^{j+1}}, r_j + \frac{\epsilon}{2^{j+1}}], \ldots\}$$

which cover the rationals and have total length ϵ. Thus the inf is clearly 0. This calculation depends very much on the rational being countable and shows that μ^* is 0 is zero for any countable set. However there are uncountable sets which get assigned 0 by μ^*; the reader is invited to check this for the Cantor set.

What properties does μ^* have? It is clearly monotone but not countably additive. We claim that it satisfies the convexity property

$$\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i)$$

where $\{A_i\}$ is a countable family not necessarily pairwise disjoint. To see this we adopt a common trick in analysis. We want to prove that some quantity, say K, is less than some other quantity, say L, both defined through a limiting process. It is often easier to prove the equivalent statement that for any ϵ, no matter how small, $K \leq L+\epsilon$; the ϵ gives one room to manoeuvre. We proceed as follows. Suppose that one does have an arbitrary positive ϵ. Since $\mu*$ is defined as the infimum of m over all covers we can find a family of intervals $\mathcal{I}_1$ covering A_1 with $m(\mathcal{I}_1) \leq \mu^*(A_1) + \frac{\epsilon}{2}$. We have written $m(\mathcal{I}_1)$ as a shorthand for the sum of the lengths of all intervals in the family $\mathcal{I}_1$. Similarly we can find a family $\mathcal{I}_n$ covering A_n with $m(\mathcal{I}_n) \leq \mu^*(A_n) + \frac{\epsilon}{2^n}$. Notice how we have used the ϵ; normally we would use the definition of μ^* as an inf to conclude that $\mu^*(A_1) \leq m(\mathcal{I}_1)$ (which is, of course, true but useless) but the extra "room" given by the epsilon allows us to get an inequality in the reverse direction. Adding up all the inequalities we get $m(\mathcal{I}) \leq \sum_i \mu^*(A_i) + \epsilon$ where the family $\mathcal{I}$ is the union of all the families $\mathcal{I}_n$. Clearly $\mathcal{I}$ covers $\cup_i A_i$ so $\mu^*(\cup_i A_i) \leq m(\mathcal{I}) \leq \sum \mu^*(A_i) + \epsilon$. Since this holds for any ϵ we have $\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i)$.

We abstract out the properties of μ^* in a definition.

Definition 2.11 An **outer measure** on a set X is a set function $\mu^* : \mathcal{P}(X) \longrightarrow [0, \infty]$, defined on all subsets of X, such that

(1) $\mu^*(\emptyset) = 0$,
(2) $A \subseteq B \Rightarrow \mu^*(A) \leq \mu^*(B)$ and
(3) for any countable family of subsets of X, say $\{A_i\}$, we have

$$\mu^*(\cup_i A_i) \leq \sum_i \mu^*(A_i).$$

We have shown that the example μ^* we have constructed on the reals is an outer measure. This is still not a measure, but it is defined on all sets. It turns out that there is a general construction which produces a σ-algebra, and a measure defined on it, from an outer measure; the σ-algebra will not, in general, be the collection of all sets. We have to choose those sets that behave nicely with respect to the outer measure when we take complements. Roughly speaking, the σ-algebra consists of those sets which can be used to "split" any other sets. Applied to our example it produces the Lebesgue measure.

Theorem 2.7 *Let X be a set and let μ^* be an outer measure defined on X. Denote by Σ the collection of all subsets, say A, such that for every subset E of X we have*

$$\mu^*(E) = \mu^*(A \cap E) + \mu^*(A^c \cap E).$$

For all A in Σ define $\mu(A) = \mu^*(A)$. Then (X, Σ, μ) is a measure space.

Proof. It is trivial to verify that $\emptyset$ is in Σ and that Σ is closed under complementation. In order to verify countable additivity we first verify finite additivity.
Suppose that $A, B \in \Sigma$. We have, for any subset E of X,

$$\mu^*(E) = \mu^*(A \cap E) + \mu^*(A^c \cap E)$$
$$= \mu^*(A \cap B \cap E) + \mu^*(A \cap B^c \cap E) + \mu^*(A^c \cap E).$$

Now we note that

$A \cap B^c = (A \cap B)^c \cap A$, hence
$(A \cap B^c) \cap E = ((A \cap B)^c \cap A) \cap E$, hence
$(A \cap B^c) \cap E = ((A \cap B)^c \cap E) \cap A.$

Similarly we have

$$A^c \cap E = ((A \cap B)^c \cap A^c) \cap E = ((A \cap B)^c \cap E) \cap A^c.$$

Using these set-theoretic identities to rewrite the last two terms in the expression for $\mu^*(E)$ we get

$$\mu^*(E) = \mu^*(A \cap B \cap E) + \mu^*(((A \cap B)^c \cap E) \cap A) + \mu^*(((A \cap B)^c \cap E) \cap A^c.$$

Using the fact that A is in Σ we can combine the last two terms

$$\mu^*(E) = \mu^*(A \cap B \cap E) + \mu^*((A \cap B)^c \cap E).$$

The last equality shows that $A \cap B$ is in Σ. Now an easy induction shows that Σ is closed under finite intersections (and hence unions too). We have shown that Σ is a field and hence a π-system.

We next show that Σ is closed under countable unions using essentially a limiting version of the finite additivity argument. Note, however, that it suffices to prove this for pairwise disjoint families because of the $\lambda - \pi$ theorem. We already know that Σ is a π-system so if we show that it is a λ-system that will establish that it is a σ-algebra.

In this case we can use the following simple identity. If A, B are disjoint sets in Σ and E is any set then

$$\mu^*((A \cup B) \cap E) = \mu^*(A \cap E) + \mu^*(B \cap E).$$

This is an easy exercise and trivially extends to any finite pairwise disjoint family. Suppose that we have a countable pairwise disjoint family $\{A_i | i \in \mathbb{N}\}$ in Σ. The union of any finite subfamily, in particular $\{A_i | i \leq n\}$, is in Σ.

Using the defining property of Σ for any set E we have
$\quad \mu^*(E) = \mu^*((\cup_{i=1}^n A_i) \cap E) + \mu^*((\cup_{i=1}^n A_i)^c \cap E)$
repeatedly using the simple identity of the last paragraph we get
$\quad = \sum_{i=1}^n \mu^*(A_i \cap E) + \mu^*((\cup_{i=1}^n A_i)^c \cap E)$.
Now we use monotonicity of μ^* to get
$\quad \mu^*(E) \geq \sum_{i=1}^n \mu^*(A_i \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E)$.
Since this is true for any $n \in \mathbb{N}$ we have
$\quad \mu^*(E) \geq \sum_{i=1}^\infty \mu^*(A_i \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E)$.
Now using subadditivity of outer measures we get
$\quad \geq \mu^*((\cup_{i=1}^\infty A_i) \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E)$
and subadditivity again gives
$\quad \geq \mu^*(E)$.

All these inequalities must therefore be equalities and we have

$$\mu^*(E) = \mu^*((\cup_{i=1}^\infty A_i) \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E)$$

which establishes that $\cup_i A_i$ is in Σ. Thus Σ is indeed a σ-algebra.

We need to show that μ, i.e. μ^* restricted to the sets in Σ, is countably additive, i.e. is a measure. In the preceding calculation we have a countable pairwise disjoint family $\{A_i\}$ of sets in Σ. We take E to be $\cup_i A_i$ in the equation

$$\mu^*(E) = \sum_{i=1}^\infty \mu^*(A_i \cap E) + \mu^*((\cup_{i=1}^\infty A_i)^c \cap E)$$

to get

$$\mu(\cup_i A_i) = \sum_{i=1}^\infty \mu(A_i)$$

where the second term drops out since it is just μ^* of the empty set. □

The theorem just proved allows one to cut down an outer measure defined on all sets to a measure defined on a σ-algebra. Sometimes one has a measure already given on a σ-algebra and wants to know whether it is unique. The next proposition says that if two measures on a σ-algebra agree on a generating π-system then they are the same. This is very useful as it is much easier to work with the π-system rather than the generated σ-algebra.

Proposition 2.10 *Suppose that $\mathcal{P}$ is a π-system on a set X and $\sigma(\mathcal{P})$ is the σ-algebra that it generates. Let P and Q be two probability measures on $\sigma(\mathcal{P})$ such that P and Q agree on all the sets in $\mathcal{P}$, then P and Q agree on all of $\sigma(\mathcal{P})$.*

Proof. We will show that the collection of sets on which P and Q agree, say Λ, form a λ-system. From this and the $\lambda - \pi$ theorem we get the result immediately because the assumption $\mathcal{P} \subset \Lambda$ implies $\sigma(\mathcal{P}) \subseteq \Lambda$.

Note first of all that $\Lambda \subseteq \sigma(\mathcal{P})$ since P and Q are only defined on sets in $\sigma(\mathcal{P})$. Since P and Q are probability measures we have $P(X) = 1 = Q(X)$ so P and Q agree on X. If $A \in \Lambda$, i.e. $P(A) = Q(A)$, then $P(A^c) = 1 - P(A) = 1 - Q(A) = Q(A^c)$ so $A^c \in \Lambda$. If $A_1, A_2 \ldots \in \Lambda$ and they are pairwise disjoint then – since P and Q are measures and all the A_i are in $\sigma(\mathcal{P})$ and pairwise disjoint –

$$P(\cup_i A_i) = \sum_i P(A_i) = \sum_i Q(A_i) = Q(\cup_i A_i).$$

Thus $\cup_i A_i$ is in Λ. This completes the proof that Λ is a λ-system. □

The last proposition requires us to know that we are working with measures. Another very useful type of theorem constructs a measure on a σ-algebra by starting with data on a restricted family of sets that generate the σ-algebra. We state a typical theorem of this kind.

Definition 2.12 A family $\mathcal{F}$ of subsets of X is called a **semi-ring** if

(1) $\emptyset \in \mathcal{F}$,
(2) $A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$ and,
(3) if $A \subseteq B$ are in $\mathcal{F}$ then there are *finitely* many *pairwise disjoint* subsets $C_1, \ldots, C_k \in \mathcal{F}$ such that $B - A = \cup_{i=1}^{k} C_i$.

This is not the form of the definition that one is used to in algebra because of the strange last condition but this is precisely the property that holds for "hyperrectangles" in $\mathbf{R}^n$.

Theorem 2.8 *Suppose that $\mathcal{F}$ is a semi-ring on X and $\mu : \mathcal{F} \longrightarrow [0, \infty]$ satisfies*

(1) $\mu(\emptyset) = 0$,

(2) μ is finitely additive and
(3) μ is countably subadditive.

Then μ extends to a measure on the σ-algebra generated by $\mathcal{F}$.

The proof of this theorem may be found in a standard text on probability and measure, for example the books by Kingman and Taylor [KT66], Billingsley [Bil95] and Ash [Ash72]. The intervals on the reals form a typical example of a semi-ring and the length function satisfies the conditions of the theorem so this extension theorem gives another construction of Lebesgue measure.

2.7 Nonmeasurable Sets

It is easy to give artificial examples of nonmeasurable sets. In this section we give a construction due to Vitali which produces a nonmeasurable subset of the reals.

Theorem 2.9 *There are Lebesgue nonmeasurable subsets of the real line.*

Proof. Let $\mathbf{R}$ stand for the real line. We define an equivalence relation on $\mathbf{R}$ as follows. We say $x \sim y$ for $x, y \in \mathbf{R}$ if $x - y$ is rational. Note that for every x in $\mathbf{R}$ there is a $y \in (0,1)$ with $x \sim y$. Let E be the set obtained by choosing from every $\sim$ equivalence class one element in $(0,1)$. We claim that such an E cannot be measurable.

Now let $E + r$ stand for the set $\{x + r : x \in E\}$ where r is a rational number. Now if r, r' are distinct rational numbers the sets $E+r$ and $E+r'$ are disjoint. Furthermore for any x in $(0,1)$ there is some rational number $r \in (-1,1)$ such that $x \in E+r$. To see this note that for any x in $(0,1)$ we have an element of E, say u, with $x \sim u$. Now let $r = x - u$, which must be a rational number. Then $x \in E+r$. Since both x and u are in $(0,1)$ we must have $r \in (-1,1)$.

Now consider the countable union

$$S = \bigcup_{r\in(-1,1)} E + r$$

where the r are all rational. The set S contains the open interval $(0,1)$ and is contained in the open interval $(-1,2)$. Furthermore all the $E+r$ are translates of E. Thus if μ is any translation-invariant measure on $\mathbf{R}$ we have $\mu(E) = \mu(E+r)$ for any r. Suppose that $\mu(E) = \alpha$ then $\mu(S) = \Sigma_{r\in(-1,1)}\alpha$

which is either 0 (if $\mu(E)$ is 0) or ∞. Now it is clear that

$$\mu((0,1)) \leq \mu(S) \leq \mu((-1,2)).$$

If in addition we want μ to agree with the lengths of intervals we have $1 \leq \mu(S) \leq 3$. But this contradicts $\mu(S) = 0$ or ∞. □

The manifest use of the axiom of choice raises the question of whether there are models of ZF which do not validate the axiom of choice and for which all subsets of the reals are measurable. One can drop the axiom of choice and construct models of ZF set theory in which the axiom of choice is false and all subsets of the reals are Lebesgue measurable. On the other hand the resulting universe has some rather unpleasant properties. For example the reals can be expressed as a countable union of countable sets. If we want a healthier model one can try to construct models satisfying the principle of dependent choice (DC) which is a little weaker than the axiom of choice. Solovay [Sol70] has indeed constructed a model of ZF set theory with DC such that all subsets of the reals are measurable. His construction is based on the assumption that there exists an inaccessible cardinal. The real point is that we are not likely to encounter nonmeasurable sets in the course of normal mathematical activity, but one has to be aware that they exist. A good discussion of these matters is in the book *Lectures in Set Theory* by Jech [Jec71].

2.8 Exercises

(1) Prove Proposition 2.1.
(2) Prove Proposition 2.2.
(3) Prove Proposition 2.3.
(4) Is there an infinite σ-algebra with only countably many members?
(5) Suppose that (X, Σ, μ) is a measure space and $A_1, A_2 \in \Sigma$. Prove that $\mu(A_1 \cup A_2) \leq \mu(A_1) + \mu(A_2)$.
(6) Suppose that $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \subseteq \ldots$ are all measurable subsets of some measure space. Prove

$$\lim_{n \to \infty} \mu(A_n) = \mu(\cup_i A_i).$$

Use this to prove a countable version of the convexity result of Part 2.
(7) Give an example (finite please!) of a set and σ-algebra and a measure such that the completion with respect to the measure can be consis-

tently extended to yet more sets.

(8) Prove that the Cantor set has Lebesgue measure 0.

(9) Prove that there are only $\aleph_1$ Borel sets but $\aleph_2$ Lebesgue measurable sets.

(10) Prove the "simple formula" used in the proof of Theorem 2.7. If A, B are disjoint sets in Σ and E is any set then

$$\mu^*((A \cup B) \cap E) = \mu^*(A \cap E) + \mu^*(B \cap E).$$

(11) Show that the measure space produced by Theorem 2.7 is complete.

(12) Suppose that X is an uncountable set and μ^* is an outer measure on X. We define μ^* by

$$\mu^*(E) = \begin{cases} 0 & \text{if } E \text{ is countable} \\ 1 & \text{otherwise.} \end{cases}$$

Describe the measurable sets and the measure.

This page intentionally left blank

Chapter 3

Integration

In this chapter we study Lebesgue integration and its relationship to the more familiar Riemann integration. In fact the two are intimately related. Before embarking on the formal study of integration it is worth recalling Lebesgue's own imagery for the difference between the two integrals; the following is taken from a lecture he gave on the origins of the Lebesgue integral.

Imagine a shopkeeper who keeps track of the money he receives in the course of a day. He could keep track of it by maintaining a running total as he receives the money or he could keep track of it by tossing the various coins and notes in bins and add up the totals later. The latter allows his to deal with money coming in much more quickly and is not affected by the money coming in an irregular fashion. The former method corresponds to Riemann integration while the latter to Lebesgue integration.

In Riemann integration we approximate the integral, say $\int_a^b f(x)dx$ by sums of the form $\sum_i f(x_i)(x_i - x_{i-1})$. We divide up the domain; the interval $[a, b]$ is partitioned by the x_i. Now this notion of integration is intuitive, has a strong constructive flavour and is the basis of all the approximation schemes that one encounters in numerical analysis. The main defect is that the function f cannot vary "too wildly". In Lesbegue integration one partitions the *range*. Now the functions can vary much more wildly, but the sets involved can be much more complicated that intervals; they can be arbitrary measurable sets.

It might seem that the whole point of Lebesgue is to integrate pathological functions. How likely are these functions to arise in practice? Actually the real reason for the predominance of the Lebesgue integral in modern analysis is that it has a much smoother theory. In particular, under quite general conditions one can interchange limits and integration, i.e. is some appropriate sense Lebesgue integration is "limit preserving" but Riemann integration is not.

In this chapter we will approach the Lebesgue integral fairly abstractly, i.e. we will not need details of Lebesgue measure in order to define it. We will also discuss the relationship between the Lebesgue integral and the Riemann integral.

3.1 The Definition of Integration

How do we divide up the range? Recall the *simple functions* which we introduced at the end of the last chapter. These are measurable functions whose range is a finite set; thus they serve to define a finite partition of the range space. They will be the functions that we start our definition with and from there we extend to the general measurable functions by a limiting process. Recall that all measurable functions are limits of sequences of simple functions.

Suppose that we have a simple function from the reals to the reals, say s, whose range is the set $\{a_1, \ldots, a_n\}$. We define $\forall i \in \{1, \ldots, n\}.A_i \overset{\mathrm{def}}{=} s^{-1}(a_i)$. The A_i are measurable sets if s is a measurable function but they could be quite complicated otherwise; far more complicated than the intervals that arise in Riemann integration. The natural definition for the integral of s is

$$\int s \,\mathrm{d}\mu = \sum_i a_i \mu(A_i)$$

where μ is Lebesgue measure. This pleasant picture is complicated by questions of well definedness immediately. What if $\mu(A_1)$ is infinite and $a_1 > 0$? It is reasonable to assign the value ∞ to the integral in this case, but what if, in addition, $\mu(A_2) = \infty$ and $a_2 < 0$? Thus it is entirely possible for a function to be measurable but not have a sensible integral.

Definition 3.1 We say that a simple function s is **integrable** if whenever a is in the range of s, $a \neq 0 \Rightarrow \mu(s^{-1}(a)) < \infty$.

It is possible to define integrable to mean that the sums arising in the definition of the integral are well-defined but the present definition is the usual one in the analysis literature. We can now make the proposed definition above official.

Definition 3.2 Suppose that (X, Σ, μ) is a measure space and that $s : X$

$\longrightarrow \mathbf{R}$ is an integrable, simple function with range $\{a_1 \ldots, a_n\}$. We say that the **integral** of s over X with respect to the measure μ is $\int_X s\mu = \sum_{i=1}^{n} a_i \mu(s^{-1}(a_i))$.

Before we move on to the definition of the integral of more general functions we need to observe some possible pathologies that have to be dealt with. Consider, for definiteness, functions from $\mathbf{R}$ to $\mathbf{R}$. We might try to make use of the fact that every measurable function is the pointwise limit of a sequence of simple functions to define the integral of general measurable functions. Now suppose that we have the family $s_i(x) = 1$ if x is between i and $i + 1$ and 0 otherwise. The limit of this family is the everywhere 0 function whose integral should certainly be 0. However the integral of every function in the family is 1. It is important therefore that we use approximation from below. Thus if we have an increasing family of simple functions $s_1 \leq \ldots \leq s_i \leq s_{i+1} \leq \ldots$ with f as the supremum we can hope to define the integral of f as the limit of the integrals of the s_i.

Now consider the function $\frac{-1}{1+x^2}$. We would like to say

$$\int_{-\infty}^{\infty} \frac{-1}{1+x^2}\, \mathrm{d}x = -\pi.$$

Unfortunately if we use our notion of approximation by simple functions *from below*, a simple function below this function could be everywhere large in absolute value and negative, and hence not integrable. Thus it is clear that we should try to approximate negative functions from above and positive functions from below. A general function has to be split into its positive and negative parts and have the integrals defined separately for each part.

For the rest of this section we fix a measure space (X, Σ, μ). When we say "real-valued function" we mean a measurable real-valued function defined on X. Suppose f is a real-valued function defined on X. We write $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$; clearly both f_+ and f_- are measurable if f is.

Definition 3.3 Suppose that f is an everywhere nonnegative real-valued function. We say that f is **integrable** if the everywhere nonnegative simple functions less than f are integrable and their integrals are bounded. If f is integrable we define

$$\int_X f\mu = \sqcup \int_X s\mu$$

where the sup is over all nonnegative simple functions below f. If we have a measurable function g which takes on both positive and negative values we say that g is integrable if both g_+ and g_- are integrable and we set

$$\int_X g\mu = \int_X g_+\mu - \int_X g_-\mu.$$

Example 3.1 We take as our measure space (X, Σ, δ_x) where δ_x is the Dirac measure concentrated at the point x of X. Let f be any nonnegative real-valued function. We claim that

$$\int_X d\delta_x = f(x).$$

Note that the simple function $s(x) = f(x)$ and 0 everywhere else is a simple function below f. The integral of s wrt δ_x is $f(x)$. Furthermore any simple function t below f has the integral $t(x) \leq f(x)$. Thus the sup of the integrals of all the simple functions below f is precisely $f(x)$.

The next example is the standard advertisement for the superior generality of the Lebesgue integral.

Example 3.2 Let $f : [0, 1] \longrightarrow \mathbf{R}$ be given by $f(x) = 0$ if x is rational and $f(x) = 1$ if x is irrational. This f is in fact a simple function, in fact it is even the characteristic function of a measurable set. Thus its integral is just the measure of the irrationals between 0 and 1 which is 1.

One has to be careful about how the word "integrable" is used. Its use suggests that a function that is not integrable cannot have a sensible integral assigned to it. The definition is rather conservative and often people would like to say that certain integrals are defined but "divergent". Thus, for example, it is common to say that $\int_{-\infty}^{\infty} 1dx = \infty$. One uses the phrase "has a definite integral" for the less stringent statement. Thus one says that the function $\lambda x.1$ is not integrable but has a definite integral between $-\infty$ and ∞ of ∞.

Example 3.3 The identity function on the reals is measurable (even continuous) but not integrable. This is a more troublesome example. A so-called pragmatic view of this is that the integral is 0; for example one can argue by symmetry. While this is what physicists and engineers usually say, the correct statement is that this function is not integrable and does not have a definite integral over the given range.

3.2 Properties of the Integral

Now we can prove some basic properties of the integral. It is customary to introduce the notation $\int_A f\mu$ for the integral of f restricted to the measurable subset A of X with the induced measure. In the next proposition functions and integrals are always on X and f, g are used for integrable functions.

Proposition 3.4

(1) If $0 \leq f \leq g$ then $\int f\mu \leq \int g\mu$.
(2) If $0 \leq f$ and $0 \leq c$ is a constant then $\int cf\mu = c\int f\mu$.
(3) $\int_A f\mu = \int_X f\chi_A\mu$ where χ_A is the characteristic function of A.

The proofs are elementary and are left to the exercises. The last is a triviality. The next proposition is the first step towards establishing linearity of the integral.

Proposition 3.5 *Let s, t be simple integrable functions. Then*

$$\int (s+t)\mu = \int s\mu + \int t\mu.$$

Proof. Let A be a measurable set, s any simple integrable function and define $\phi(A) = \int_A s\mu$. Then ϕ is a measure. To see this recall that any simple function can be written in the form of a finite sum

$$s = \sum_{i=1}^{n} a_i \chi_{A_i}$$

where for definiteness we assume that the a_i are distinct. Now suppose that $C = \cup_{i\in\mathbf{N}} C_i$ where the C_i are pairwise disjoint and measurable. We have

$$\begin{aligned}\phi(C) &= \textstyle\sum_{i=1}^{n} a_i\mu(A_i \cap C)\\ &= \textstyle\sum_{i=1}^{n} a_i \sum_{j=1}^{\infty} \mu(A_i \cap C_j)\\ &= \textstyle\sum_{j=1}^{\infty} \sum_{i=1}^{n} a_i\mu(A_i \cap C_j)\\ &= \textstyle\sum_{j=1}^{\infty} \phi(C_j)\end{aligned}$$

where the first equality is by definition of ϕ, the second by countable additivity, the third by absolute convergence and the fourth by definition. Thus we have shown that ϕ is countably additive.

Let t be written as $t = \sum_{j=1}^{m} b_j\chi_{B_j}$ where the b_j are all distinct. Let $C_{ij} = A_i \cap B_j$. The sets C_{ij} form a partition of X. Now $s+t$ is simple and integrable. Thus $\psi(D) = \int_D (s+t)\mu$ is a measure. Thus we can write

$$\begin{aligned}\textstyle\int_X (s+t)\mu &= \textstyle\sum_{i,j}\int_{C_{ij}}(s+t)\mu\\ &= \textstyle\sum_{i,j}(a_i+b_j)\mu(C_{ij})\\ &\textstyle\sum_{i,j}(\int_{C_{ij}} s\mu + \int_{C_{ij}} t\mu)\\ &= \textstyle(\sum_{i,j}\int_{C_{ij}} s\mu) + (\sum_{i,j}\int_{C_{ij}} t\mu)\\ &= \textstyle\int_X s\mu + \int_X t\mu\end{aligned}$$

□

We are now ready to prove one of the most important theorems in integration theory and the most useful for our purposes. This theorem is Lebesgue's monotone convergence theorem and represents perhaps the most striking difference between the Riemann integral and the Lebesgue integral. Roughly speaking it says that the integral of the limit of a sequence of functions is the limit of the integrals of the individual functions. There are several theorems like this and some, for example the dominated convergence theorem, is more general and more useful in classical analysis, but the starting point of all such theorems is the monotone convergence theorem.

Theorem 3.6 ***[Monotone convergence theorem]*** *Let $\{f_n\}$ be a sequence of measurable functions on X and suppose that*

(1) $\forall x \in X. 0 \leq f_1(x) \leq f_2(x) \leq \ldots \leq \infty$
(2) $\forall x \in X. \sqcup_n f_n(x) = f(x)$.

Then f is measurable, and $\sqcup_n \int_X f_n\mu = \int_X f\mu$.

Of course we already know that f is measurable, it is only asserted here to ensure that the subsequent statement is sensible. Usually one states this with limits rather than sups, the choice is a matter of taste. We work with the complete lattice $[0,\infty]$ so that we might be dealing with functions that are not integrable but do possess a definite integral. For everywhere nonnegative functions, having a definite integral is the same as being measurable.

Proof. We know that $\forall n. \int_X f_n\mu \leq \int_X f\mu$ so $K \overset{\text{def}}{=} \sqcup_n \int_X f_n\mu \leq \int_X f\mu$. We need to prove that $K \geq \int f\mu$. The latter is also a sup, $\int_X f\mu = \sqcup \int_X s\mu$ where s stands for a simple function below f. Thus if we can prove that for any simple function $s \leq f$ $\int_X s\mu \leq K$ we will be done.

What we shall prove is that for any number c in the interval $(0,1)$ we have $c\int_X s\mu \leq K$ from which the required inequality immediately follows. Accordingly we fix such a c and a simple function s. We define the sets

$$\forall n. A_n = \{x | f_n(x) \geq c \cdot s(x)\}.$$

These are clearly measurable sets and $\forall n. A_n \subseteq A_{n+1}$ since the f_n form an increasing sequence. Now we claim that $\cup_n A_n = X$. Let $x \in X$ be arbitrary. Since $c < 1$ we have $c \cdot s(x) < f(x)$ if $f(x) > 0$. Since $\sqcup_n f_n(x) = f(x)$ there is some m such that $f_m(x) > c \cdot s(x)$ and hence $x \in A_m$. If $f(x) = 0$ then clearly $c \cdot s(x) = 0$ and $x \in A_1$. Now we clearly have

$$\int_X f_n \mu \geq \int_{A_n} f_n \mu \geq c \int_X s\mu.$$

If we take sups of both sides on the left we get K and on the right we have that $\cup_n A_n = X$, thus

$$K \geq c \cdot \int_X s\mu.$$

□

It is worth noting how the constant c is used to make the A_n big enough to cover X. We have the immediate corollary

Corollary 3.7 *If we have f and f_n as above and A is any measurable subset then $\int_B f\mu = \sqcup \int_B s\mu$.*

The next theorem is a basic useful fact but the real point of discussing it now is to exhibit a paradigmatic proof strategy for using the monotone convergence theorem. Let $T : (X, \Sigma_X) \longrightarrow (Y, \Sigma_Y)$ be a measurable function. A measure μ on X induces a measure ν on Y by the formula $\nu = \mu \circ T^{-1}$. If $X = Y$, $\Sigma_X = \Sigma_Y$ and $\mu = \nu$ we call T an *ergodic* transformation.

Proposition 3.8 *If we have X, Y, T as above and $f : Y \longrightarrow \mathbf{R}$ is measurable and $B \in \Sigma_Y$ then*

$$\int_{T^{-1}(B)} f \circ T\mu = \int_B f\nu$$

in the sense that if one integral exists then so does the other and they are equal.

Proof. Suppose that f is the characteristic function of some measurable set C. Then $f \circ T$ is the characteristic function of the set $T^{-1}(C)$. The left hand side of the above equation is then $\mu(T^{-1}(C) \cap T^{-1}(B))$. But this is equal to $\mu(T^{-1}(B \cap C)) = \nu(B \cap C)$ which is precisely the value of the right hand side. Since the integral is linear on simple functions and any simple function is a linear combination of characteristic functions, the required equation holds for any simple function. Now consider any nonnegative function f. We have $f = \sqcup\{s | 0 \leq s \leq f\}$. The monotone convergence theorem says that $\int_B f\nu = \sqcup \int_B s\nu$, while the fact that s is simple allows us to conclude that $\int_B s\nu = \int_{T^{-1}(B)} s \circ T\mu$. Now we also note that $\sqcup s \circ T = f \circ T$. Putting all these together we get the required result. For f not nonnegative we can easily apply these arguments separately to the positive and negative parts. □

This proof is paradigmatic. In fact this pattern is so closely followed that one can omit the details and just invoke the pattern of this proof as a kind of "mantra". When I say "monotone convergence mantra" henceforth I mean exactly this pattern. A few exercises will familiarise any readers with this proof and then they will be able to recognise occurrences of this pattern easily. A typical application is to prove that integration is indeed linear for integrable functions, not just for simple functions.

Proposition 3.9 *If f and g are measurable functions then*

$$\int (f+g)\mu = \int f\mu + \int g\mu.$$

The proof is left as an exercise.

The most important theorem relating limits and integration is Lebesgue's dominated convergence theorem. We state this without proof.

Theorem 3.10 *Let f_n and g be integrable functions such that*

$$\forall n. |f_n| \leq g \text{ and } \lim_{n \longrightarrow \infty} f_n = f.$$

Then f is integrable and

$$\lim_{n \longrightarrow \infty} \int f_n\mu = \int f\mu.$$

This allows one to deal with sequences of functions that are not monotone increasing. However this theorem follows essentially from the monotone convergence theorem.

3.3 Riemann Integrals

In this section we discuss the relationship between the familiar Riemann integral and the Lebesgue integral. Consider real-valued functions defined on a closed interval $[a, b]$.

Definition 3.11 We say that such a function f is Riemann-integrable with Riemann integral α if

$$\forall\epsilon > 0 \exists a = a_0 < a_1 < a_2 < \ldots < a_n = b.|(\sum_{i=1}^{n} f(x_i)(a_i - a_{i-1}) - \alpha| \leq \epsilon$$

whenever $x_i \in [a_{i-1}, a_i]$.

Note the way the definition is couched in terms of dividing up the domain of the function.

Example 3.4 Suppose that the interval is $[0, 1]$ and the function is $f(x) = x$. To verify that the integral is $\frac{1}{2}$ we proceed as follows. Given an ϵ we choose the $a_i = \frac{i}{n}$, so $a_i - a_{i-1} = \frac{1}{n}$. We maximise the sum by choosing $x_i = a_i$ and we minimise it by choosing $x_i = a_{i-1}$. Easy calculations then show that

$$\sum_{i=1}^{n} f(x_i)(a_i - a_{i-1}) \in [\frac{1}{2} - \frac{1}{2n}, \frac{1}{2} + \frac{1}{2n}].$$

In other words by choosing $n > \frac{2}{\epsilon}$ and dividing up the interval uniformly we get the required inequality with $\alpha = \frac{1}{2}$.

The next example is a contrast with the Lebesgue case.

Example 3.5 We again consider a function defined on $[0, 1]$. Suppose that f is 0 if x is rational and 1 if x is irrational. The Lebesgue integral of this function is 1; we shall argue that the Riemann integral does not exist. Let us choose $\epsilon = \frac{1}{2}$. Let the a_i be chosen arbitrarily. Any of the intervals $[a_{i-1}, a_i]$ will contain both rational and irrational numbers. Choosing only rationals we see that the sum is zero while choosing only irrationals the sum is 1. Thus whatever α might be there is no way of forcing the sum to be within ϵ of α for all choices of the x_i.

This example shows the role played by continuity. If f is too "wild" we cannot constrain the values of the sum by confining x to small intervals.

Theorem 3.12 *Let f be a bounded function on $[0, 1]$. The function f is Riemann-integrable iff the set of points at which f is not continuous has Lebesgue measure 0.*

Proof. We first define a function F which "measures how discontinuous f is." For a fixed $x \in [0, 1]$ we define

$$F(x) = \lim_{\epsilon \longrightarrow 0} sup_{x', x''}\{|f(x') - f(x'')| : |x - x'|, |x - x''| \leq \epsilon\}.$$

The sup is defined since f is bounded and the sup decreases as $\epsilon \longrightarrow 0$ so the limit is defined. Clearly if f is continuous at x, $F(x) = 0$. At a point of discontinuity of f, F is the value of the jump in the value of f, well-defined since f is bounded. We define $F_\delta = \{x : F(x) \geq \delta\}$. The set of all points where f is discontinuous is $D = \cup_\delta F_\delta$.

Let us assume that f is Riemann-integrable. We claim that for all δ the set F_δ has Lebesgue measure 0. To see this we fix a δ. Now for any $\epsilon > 0$ we can find a_i as required by the definition of Riemann-integrable. Now suppose that I is the subset of $\{1, \ldots, n\}$ such that that every interval $[a_{i-1}, a_i]$ with $i \in I$, contains a point of F_δ, then the entire sum in the definition of the Riemann integral can vary by at least $\delta \cdot \sum_{i \in I}(a_i - a_{i-1})$. But since f is Riemann-integrable we must have this sum bounded by $2 \cdot \epsilon$. Thus $\sum_{i \in I}(a_i - a_{i-1}) \leq \frac{2\epsilon}{\delta}$. Now F_δ, except for finitely many points (some of the a_i may be in F_δ), is contained in intervals of total length $\frac{2\epsilon}{\delta}$. Thus the outer measure of F_δ is less than $\frac{2\epsilon}{\delta}$, which is only possible for all ϵ if the outer measure, and hence the Lebesgue measure, of F_δ is 0.

Now D is contained in the union of countably many sets of measure 0

$$D \subseteq \bigcup_j F_{\frac{1}{2^j}},$$

and thus has measure 0.

To argue the other direction we use compactness. Essentially we have to show that we can construct a partition such that we can make the variation of the sum appearing in the definition of the Riemann integral as small as we please.

Recall that closed and bounded subsets of $\mathbf{R}$ are compact. We claim that F_δ is always closed. Suppose that we have a convergent sequence $\{x_i\}$ in F_δ and that the sequence converges to x. We need to argue that x is in F_δ. Any open neighbourhood of x contains some point x_i, i.e. some point of F_δ, so in any neighbourhood of x, the variation of f is at least δ. Thus $x \in F_\delta$.

Since by assumption F_δ has measure 0, it can be covered by a family of open intervals of total length less than any specified ϵ. Since F_δ is compact, a finite number of these intervals suffice to cover F_δ. Without loss of generality we can take these intervals to be disjoint since the union of two overlapping open intervals is an open interval. We call these intervals $A_1 \ldots, A_n$. We define $B = (\cup_i A_i)^c$; clearly B is closed hence compact. Also F takes on values less than δ on B. We can therefore find finitely many open, disjoint intervals that cover B; we call these B intervals.

We construct a partition which bounds the variation in the sum appearing in the Riemann integral as follows. We order the end points of the A and the B intervals in order to define the partition. The intervals may overlap and we may overestimate the variation but we will bound this estimate in the right direction. The variation in the value of the sum just taken over the A intervals is $2M\epsilon$ (where M is the bound on the value of f) since their total length is bounded by ϵ. The contribution to the variation of the sum coming from the B intervals is δ. The total length of the B intervals may be 1 but the variation cannot be more than δ. Thus the overall bound we get is $\delta + 2M\epsilon$. Since we can choose both δ and ϵ to be arbitrarily small we can make the variation in the sum appearing in the definition of the Riemann integral arbitrarily small. □

The Riemann integral and the Lebesgue integral are intimately related. As the above theorem shows Riemann-integrable functions are characterised by Lebesgue measure properties of the discontinuities. A more intimate tie-up is revealed by the Riesz representation theorem. This very important theorem is beyond the scope of these notes but we can outline the ideas. The space of continuous functions on a closed interval forms a topological vector space[1]. The vector space structure is evident, the topological structure comes from the sup norm. Now Riesz's theorem says that any *continuous, linear* functional, say Λ, on this space can be represented by a measure μ, in other words one can find a measure μ such that

$$\Lambda(f) = \int f\mu.$$

For example the functional that evaluates a function at a point is represented by the Dirac measure. All continuous functions are Riemann-integrable so the Riemann integral defines a continuous linear functional. The measure that corresponds to this functional is precisely Lebesgue measure. In many books Lebesgue measure is first constructed this way.

3.4 Multiple Integrals

A very basic result in integration is the ability to treat double integrals as iterated integrals and to change the order of integration of the latter. The basic result, which says that under very general conditions one can do all

[1] It also forms an algebra and various other things, but we concentrate on the TVS structure for now.

this, is called Fubini's theorem. In order to treat this we need a little more measure theory. We see that some of these basic facts are proved by using integration, in particular the monotone convergence theorem.

We first need to define measures on product spaces. Given two measure spaces (X, Σ_X, μ) and (Y, Σ_Y, ν) we want to define a measure on $(X \times Y, \Sigma_X \otimes \Sigma_Y)$. We call a set of the form $A \times B$, with $A \in \Sigma_X, B \in \Sigma_Y$, a *measurable rectangle*. We can define a measure ρ on $\Sigma_X \otimes \Sigma_Y$ by starting with $\rho(A \times B) = \mu(A) \cdot \nu(B)$. We use our usual conventions for arithmetic with ∞. Now in order to define a measure we need to know that ρ is countably additive.

Proposition 3.13 *ρ is countably additive on the collection of rectangles.*

Proof. Suppose that we have a countable pairwise disjoint family of rectangles $A_n \times B_n$ such that there exist A and B with $A \times B \cup_n A_n \times B_n$.

We work with characteristic functions, thus we have

$$\chi_A \chi_B = \sum_n \chi_{A_n} \chi_{B_n}.$$

Now for fixed x we have on each side a measurable function of y. Integrating wrt y using the measure ν gives the equation

$$\chi_A(x)\nu(B) = \sum_n \chi_{A_n}(x)\nu(B_n).$$

This is immediate from countable additivity of ν. Now we can integrate with respect to x using μ to get

$$\mu(A)\nu(B) = \sum_n \mu(A_n)\nu(B_n)$$

where we have used linearity of the integral and the monotone convergence theorem in the usual way. □

The rectangles form a semi-ring thus Theorem 2.8 says that ρ extends to a measure on the σ-field generated by the rectangles. However, this extension is not unique. The key property needed to guarantee uniqueness is σ-finiteness ; i.e. the measure space is the countable union of sets of finite measure. In probability theory this condition is usually satisfied.

We are now ready to show that we have a *unique* product measure when the measures in question are σ-finite. It turns out useful to work with monotone classes. We follow the general pattern of the monotone convergence mantra and begin by showing that we can invert the order of integration for characteristic functions.

Proposition 3.14 *Suppose that (X, Σ_x, μ) and (Y, Σ_Y, ν) are measure spaces such that $\mu(X) < \infty$ and $\nu(Y) < \infty$. Let*

$$\mathcal{F} \stackrel{\text{def}}{=} \{E \subseteq X \times Y : \int [\int \chi_E(x, y)\mu]\nu = \int [\int \chi_E(x, y)\nu]\mu\}.$$

Then $\Sigma_x \otimes \Sigma_Y \subseteq \mathcal{F}$.

Proof. We start with rectangles and then use the fact that the smallest monotone class containing all the rectangles is just the same as the σ-field generated by the rectangles. Let $E = A \times B$ be a rectangle. Now we have $\chi_E = \chi_A \cdot \chi_B$ so

$$\int [\int \chi_E \mu]\nu = \int [\int \chi_A \chi_B \mu]\nu = \mu(A)\nu(B) = \int [\int \chi_E \nu]\mu.$$

Thus any rectangle is in $\mathcal{F}$. Clearly any finite disjoint union of rectangles is also in $\mathcal{F}$.

Now in order to show that $\mathcal{F}$ is closed under increasing and decreasing monotone limits we proceed by induction on the stages in the construction of $\mathcal{F}$ starting from the rectangles. We have just established the base case. For the inductive case we suppose that $E_n \uparrow E$ in $\mathcal{F}$ and that we can reverse the order of integration for the characteristic functions of the E_n. Now if $E_n \uparrow E$ we have $\sqcup_n \chi_{E_n} = \chi_E$. Thus by using the monotone convergence theorem we have $\int \int \chi_{E_n} \mu\nu = \int \int \chi_E \mu\nu$ and also $\int \int \chi_{E_n} \nu\mu = \int \int \chi_E \nu\mu$. By assumption $\int \int \chi_{E_n} \mu\nu = \int \int \chi_{E_n} \nu\mu$ so we get $\int \int \chi_E \mu\nu = \int \int \chi_E \nu\mu$. Thus $E \in \mathcal{F}$. If we have $E_n \downarrow E$ we proceed in exactly the same way except that we use the dominated convergence theorem rather than the monotone convergence theorem. It is here that we need the finiteness assumption on the measure. Thus $\mathcal{F}$ is a monotone class. Note that we are always dealing with measurable functions at every stage of this proof. □

Theorem 3.15 *Let (X, Σ_x, μ) and (Y, Σ_Y, ν) be σ-finite measure spaces. Then ρ extends uniquely to a measure on $\Sigma_x \otimes \Sigma_Y$ such that for all $E \in \Sigma_x \otimes \Sigma_Y$*

$$\rho(E) = \int [\int \chi_E(x, y)\mu]\nu = \int [\int \chi_E(x, y)\nu]\mu.$$

Proof. We start by assuming that the measures μ and ν are finite. Let $\eta(E) = \int [\int \chi_E(x, y)\mu]\nu$ for $E \in \Sigma_X \otimes \Sigma_Y$. Now this integral is defined since the measures are finite, and, as we have seen, the order of integration can be reversed. The set function η is finitely additive and by using monotone convergence theorem it is also countably additive; thus it is a measure.

Suppose that we have another measure ξ with the same properties. By the same arguments as before we can prove that the sets for which ξ and η agree include all finite unions of disjoint rectangles and form a monotone class and hence include all of $\Sigma_X \otimes \Sigma_Y$. Thus we have uniqueness.

Now we extend the result to the σ-finite case. Let X_n and Y_m satisfy, $\cup_n X_n = X$, $\forall n.\mu(X_n) < \infty$, $\cup_m Y_m = Y$ and $\forall m.\nu(Y_m) < \infty$. Let $E \in \Sigma_X \otimes \Sigma_Y$ and define sets $E_{n,m}$ by $E_{n,m} = E \cap (X_n \times Y_m)$. Now for these sets we have

$$\int\int \chi_{E_{n,m}} \mu\nu = \int\int \chi_{E_{n,m}} \nu\mu.$$

We can sum all these terms and use monotone convergence to conclude that

$$\int\int \chi_E \mu\nu = \int\int \chi_E \nu\mu.$$

Thus we can define η as before. To establish that it is a measure we use countable additivity of μ and ν to show that η is countably additive. Uniqueness follows from the fact that any other measure must agree on the $E_{n,m}$ according to the previous paragraph and hence on E by monotone convergence. □

The measure whose existence is established above is called the product measure and is written $\mu \times \nu$. What we have done so far is the first step in the chain of reasoning that one needs to invoke the monotone convergence theorem. In other words we have shown that we can change the order of integration if we are integrating characteristic functions. We can now complete the argument in the familiar way and prove

Theorem 3.16 ***[Fubini-Tonelli]*** *Let (X, Σ_x, μ) and (Y, Σ_Y, ν) be σ-finite measure spaces and let f be an measurable function with respect to $\Sigma_X \otimes \Sigma_Y$ from $X \times Y$ to $[0, \infty]$. Then*

$$\int f(\mu \times \nu) = \int [\int f\mu]\nu = \int [\int f\nu]\mu.$$

The same holds for not necessarily positive integrable functions.

The integral $\int f\mu$ is defined for ν-almost all y and the integral $\int f\nu$ is defined for μ-almost all x. This will not affect the values of the integrals.

Proof. For characteristic functions this is just the last theorem. For simple functions we have the result by linearity of integration. For arbitrary positive f it follows from the monotone convergence theorem and the usual case analysis for the positive and negative parts gives the result for the case of integrable functions. □

We have stated these theorems for binary products but by an easy induction on n it holds for products of n σ-finite spaces as well. Constructing measures on larger products is more subtle but is important for the analysis of stochastic processes.

3.5 Exercises

Unless otherwise stated these exercises refer to a fixed measure space (X, Σ, μ) and the phrase "measurable function", without further qualification means "measurable real-valued function."

(1) Show that the integral is a linear function of measurable functions, i.e. for f, g measurable show that

$$\int (f+g)\mu = \int f\mu + \int g\mu.$$

(2) Show that if f_n is a family of nonnegative, real-valued, measurable functions then $f \stackrel{\text{def}}{=} \sum_n f_n$ is measurable and that

$$\int f\mu = \lim_{n \longrightarrow \infty} \int f_n \mu.$$

(3) Let A_1 be $[0,1]$, $A_2 = A_1 \cap (\frac{1}{3}, \frac{2}{3})^c$, $A_3 = A_2 \cap (\frac{1}{9}, \frac{2}{9})^c \cap (\frac{7}{9}, \frac{8}{9})^c$ and so on. We have $A \stackrel{\text{def}}{=} \cap_i A_i$; this is known as the Cantor set. Now we define a function $f : [0,1] \longrightarrow [0,1]$ by

$$f(x) = \begin{cases} 1 & \text{if } x \in A \\ \frac{1}{2} & \text{if } x \in A_1 \cap A_2^c \\ \frac{1}{4} & \text{if } x \in A_2 \cap A_3^c \end{cases}$$

Show that this function is Riemann-integrable and calculate its Riemann integral. [Hint: the answer is $\frac{1}{4}$.]

(4) The monotone convergence theorem is not true for arbitrary directed sets of measurable functions. Construct an example showing this. [Hint: Consider the interval $[0,1]$ and the characteristic functions of all finite subsets of this interval.] (I am indebted to Jørgen Hoffman-Jørensen for pointing this out to me.)

This page intentionally left blank

Chapter 4

The Radon-Nikodym Theorem

Consider the discrete notion of conditional probability. Supposing we have some probability space (X, Σ, Pr) we write

$$Pr(A|B) \overset{\text{def}}{=} \frac{Pr(A \cap B)}{Pr(B)}$$

where the lhs is read "the probability of A *given* B" and it is assumed that $Pr(B) \neq 0$. For discrete situations the above definition is adequate, though often counter-intuitive, but for continuous probability spaces one may well be dealing with situations where both the numerator and the denominator are 0. A typical example is when the set X is a product, say $\mathbf{R} \times \mathbf{R}$, with some probability density assigned to the measurable sets of the product σ-field. We might be interested in conditional probabilities of the form $Pr(A|\{y\})$ where y is a point and A is a subset of $\mathbf{R}$. Now we can use the general formula for conditional probability and get

$$Pr(A|y) = \frac{Pr(A \times \{y\})}{Pr(\mathbf{R} \times \{y\})}$$

where we are writing $Pr(A|y)$ rather than $Pr(A|\{y\})$. Unfortunately this expression rarely, if ever, makes sense; the denominator is usually 0.

Nevertheless, the notion of conditional probability is useful in this regime and it is important to try and formulate what it might mean. A heuristic account [Fel71] might proceed as follows. We imagine a nested family of measurable sets $B_1 \supseteq B_2 \supseteq \ldots \supseteq B_n \supseteq \ldots$ such that $\cap_i B_i = \{y\}$. Now we can try to use a limiting version of the discrete formula for conditional probability:

$$Pr(A|y) = \lim_{i \longrightarrow \infty} \frac{Pr(A \times B_i)}{Pr(\mathbf{R} \times B_i)}.$$

This is a reasonable intuition but making precise when the limits actually exist – and under what conditions this is a rigorous definition – are difficult problems.

What one can observe is that one needs some notion of "differentiation" of measures. The Radon-Nikodym theorem serves precisely this role. To arrive at a plausible statement we can proceed as follows. Let us think about measures defined on the reals. Consider the function $F(x)$ defined by the integral $F(x) = \int_{\infty}^{x} f(x)dx$, where dx refers to Lebesgue measure. If f has just a few finite jumps then F is well-behaved. If F has a finite jump f will have a singularity. Now a typical singularity can be thought of a Dirac delta "function", which we know can be rigorously defined as a measure concentrated at a point. Now we might think of measures which assign nonzero weight to a single point as being singular but those which do not should be essentially given by a formula like the above for F. More precisely one might conjecture

> If λ is a measure on $\mathbf{R}$ that has the property $\lambda(\{x\}) = 0$ for any x, then there is some measurable function f such that
>
> $$\lambda(B) = \int_B f(x)dx$$
>
> for any measurable set B.

Lebesgue showed that this is false, but if the hypothesis is strengthened to $\lambda(B) = 0$ whenever B has Lebesgue measure 0 it is true. The Radon-Nikodym theorem generalizes this to the abstract setting. This is precisely the notion of differentiation we need to make sense of conditional probability, as we shall see later.

4.1 Set Functions

In order to proceed we need to develop some of the theory of set functions. Much of this can be done in analogy with measures except that we need more care with convergence since we no longer assume positivity.

Definition 4.1 If (X, Σ) is a measurable space and $\lambda : \Sigma \longrightarrow [-\infty, \infty]$ is an extended real-valued function we say that λ is **countably additive** (**finitely additive**) if, whenever $\{A_i\}_{i \in I}$ is a countable (finite) family of

Σ-measurable sets that are pairwise disjoint, then

$$\lambda(\cup_{i \in I} A_i) = \Sigma_{i \in I} \lambda(A_i).$$

The following easy facts are proved just like the analogous results for measures.

Lemma 4.1 *Let λ be a countably additive set function on (X, Σ). All the sets mentioned in this lemma are assumed Σ-measurable.*

(1) If $A_i \uparrow A$, then $\lim_{i \longrightarrow \infty} \lambda(A_i) = \lambda(A)$.
(2) If $A_i \downarrow A$ and $\lambda(A_1)$ is finite, then $\lim_{i \longrightarrow \infty} \lambda(A_i) = \lambda(A)$.

The proof is left as an exercise.

4.2 Decomposition Theorems

Even though set functions are not measures they can be decomposed into positive and negative pieces. This result, the Hahn-Jordan decomposition theorem is a key step in the proof of the Radon-Nikodym theorem. The next theorem contains all the intricate set-theoretic combinatorics needed to prove the decomposition theorem. The proof closely follows the proof in Ash [Ash72].

Theorem 4.2 *Let λ be a countably additive set function on (X, Σ). Then there are sets $C, D \in \Sigma$ such that $\lambda(C) = sup\{\lambda(A) : A \in \Sigma\}$ and $\lambda(D) = inf\{\lambda(A) : A \in \Sigma\}$.*

Remark: This theorem says that λ actually attains a maximum and a minimum value. This is obviously trivial for a measure.

Proof. If we show that C exists then we can apply the result to $-\lambda$ to show that D exists, so we only need construct C. Furthermore, we can assume that λ is always finite, for if $\lambda(A) = \infty$ for some A we can take $C = A$.

Let $\alpha = sup\{\lambda(A) : A \in \Sigma\}$. There is a family of sets $A_n \in \Sigma$ with $\lim_{i \longrightarrow \infty} \lambda(A_i) = \alpha$. We need to construct a set C such that $\lambda(C) = \alpha$. Let $A = \cup_i A_i$. Now, of course, A is not the set we are trying to construct since it may contain negative pieces. In order to remove these pieces we need to carve A into sufficiently refined subsets. We can do this as the limit of successively refined decompositions of A.

We proceed as follows. We write A'_i for $A \setminus A_i$ and we write A^*_i for *either* A_i *or* A'_i. Now for each positive integer n we can construct a partition of A into 2^n pairwise disjoint sets (some of which may be empty) as follows. We take all combinations of the form

$$A^*_1 \cap A^*_2 \ldots \cap A^*_i \cap \ldots \cap A^*_n.$$

We call these sets A_{nm} where m takes on the values from 1 to 2^n. These sets are clearly pairwise disjoint and cover A. Furthermore each A_n is a union of some of the A_{nm}. If we consider $n' > n$ then each $A_{n'm'}$ is contained in some A_{nm} or is disjoint from it; we call this the refinement property. Thus we have a family of increasingly refined partitions of A.

We extract the positive pieces of A by proceeding in stages. Let B_n be defined by

$$B_n = \bigcup_m \{A_{nm} : \lambda(A_{nm}) \geq 0\}$$

where we set $B_n = \emptyset$ if there are no A_{nm} with a positive λ value. Now since each A_n is a union of some of the A_{nm} we have $\lambda(A_n) \leq \lambda(B_n)$. Note that the sequence B_n need not be monotonically increasing. As we refine, we may lose negative pieces that we used to include and also include new positive pieces. Thus we cannot just take the limit of the $\lambda(B_n)$.

What we can do, however, is take the *lim sup* of the B_n.

$$C = \limsup B_n = \bigcap_{n=1}^{\infty} (\bigcup_{k=n}^{\infty} B_k).$$

Clearly, $\bigcup_{k=n}^{\infty} B_k \downarrow C$. Now because of the refinement property of the A_{nm} we have that $\bigcup_{k=n}^{j} B_k$ can be written as the union of B_n and additional $A_{n'm'}$ with $n' > n$. These $A_{n'm'}$ will either already be in B_n or disjoint from it and they will satisfy $\lambda(A_{n'm'}) \geq 0$. Thus $\bigcup_{k=n}^{j} B_k$ can be written as the union of B_n and a set disjoint from B_n. Thus $(\bigcup_{k=n}^{j} B_k) \uparrow (\bigcup_{k=n}^{\infty} B_k)$ and hence, by Lemma 4.1, part (a), we have

$$\lambda(A_n) \leq \lambda(B_n) \leq \lambda((\bigcup_{k=n}^{j} B_k)) \longrightarrow \lambda(\bigcup_{k=n}^{\infty} B_k).$$

Recalling that $\bigcup_{k=n}^{\infty} B_k \downarrow C$ and using Lemma 4.1, part (b), we get

$$\alpha = \lim_{n \longrightarrow \infty} \lambda(A_n) \leq \lim_{n \longrightarrow \infty} \lambda(\bigcup_{k=n}^{\infty} B_k) = \lambda(C) \leq \alpha.$$

Thus $\lambda(C) = \alpha$. □

We are now able to easily prove the Hahn-Jordan theorem.

Definition 4.3 Let λ be a set function defined on the measurable space (X, Σ). A set A is called **positive** with respect to λ if $\forall E \subseteq A . \lambda(E) \geq 0$; we define **negative** sets similarly.

Theorem 4.4 ***[Hahn decomposition]*** *For any countably additive set function λ defined on the measurable space (X, Σ), there are disjoint sets A^+ and A^- such that $X = A^+ \cup A^-$ where A^+ $(A^-)$is positive (negative) with respect to λ.*

Proof. Note that λ cannot take on both the values ∞ and $-\infty$, if it did it would not be countably additive. Without loss of generality, we can assume that $-\infty$ is never attained. Let A^+ be the set on which λ attains its sup and define A^- as the complement of A^+. Let E be a measurable subset of A^+, then if $\lambda(E) < 0$ we have $\lambda(A^+ \setminus E) > \lambda(A)$, which contradicts the maximality of $\lambda(A^+)$. Thus A^+ is positive. If $E \subseteq A^-$ and $\lambda(E) > 0$ then $\lambda(A^+ \cup E) > \lambda(A^+)$ which again contradicts the maximality of E. Thus A^- is negative. □

Corollary 4.5 ***[Jordan decomposition]*** *If λ is countably additive set function defined on a measurable space (X, Σ), then there are two positive measures λ^+ and λ^- such that for all $S \in \Sigma$ we have*

$$\lambda(S) = \lambda^+(S) - \lambda^-(S).$$

Proof. Take the Hahn decomposition, A^+, A^- as above. Then we define $\lambda^+(S)$ to be $\lambda(S \cap A^+)$ and $\lambda^-(S)$ to be $\lambda(S \cap A^-)$. It is easy to verify that these are measures given that λ is countably additive. □

4.3 Absolute Continuity

We study the relation of absolute continuity between measures, as this is the key assumption in the Radon-Nikodym theorem. There are actually two closely related concepts.

Definition 4.6 Two measures, μ, ν on a measurable space (X, Σ) are **mutually singular**, written $\mu \perp \nu$, if there are disjoint measurable sets A, B with $\mu(X \setminus A) = 0$ and $\nu(X \setminus B) = 0$.

Definition 4.7 Suppose that μ and ν are measures defined on a measurable space (X, Σ). We say that ν is **absolutely continuous** with respect to μ, written $\nu << \mu$, if $\forall A \in \Sigma.\mu(A) = 0 \implies \nu(A) = 0$.

Clearly if f is a measurable function and we define ν by $\nu(A) = \int_A f\mu$ we get a measure such that $\nu << \mu$. The Radon-Nikodym theorem essentially goes in the opposite direction.

Theorem 4.8 ***[Radon-Nikodym-Lebesgue]*** *If μ and ν are both σ-finite measures on a measurable space (X, Σ) then:*

(1) ν can be written as $\nu_a + \nu_s$ where $\nu_a << \mu$ and $\nu_s \perp \mu$.
(2) There is a non-negative measurable function f such that

$$\forall A \in \Sigma. \int_A f\mu = \nu_a(A).$$

If g is another function satisfying the same property as f then the set of points where f and g differ have μ measure 0.

Part (1) of the theorem is usually called the Lebesgue decomposition while part (2) is usually called the Radon-Nikodym theorem. The function f (unique μ-almost everywhere) is often called the Radon-Nikodym derivative and is written $\frac{d\nu_a}{d\mu}$.

Proof. If part (2) of the theorem is true for the case of finite measures it extends to σ-finite measures as follows. We assume that $\nu << \mu$. Since μ is assumed σ-finite we have a countable family of pairwise disjoint sets A_n which cover X and satisfy $\mu(A_n) < \infty$. We define $\mu_n(A) = \mu(A \cap A_n)$ and similarly for ν_n. Now μ_n, ν_n are finite measures and, by assumption, we have $\nu_n << \mu_n$ whence there is a measurable function f_n such that $\forall A \in \Sigma.\nu_n(A) = \int_A f_n\mu_n$. Now we know that

$$\nu(A) = \sum_n \nu_n(A) = \sum_n \int_A f_n\mu_n = \sum_n \int_A f_n\chi_{A_n}\mu = \int_A \sum_n f_n\chi_{A_n}\mu.$$

The last equality follows from the monotone convergence theorem. Thus the required f is $\sum_n f_n\chi_{A_n}$.

Thus for part (2) it suffices to establish the result for finite measures. We will discuss part (1) later. We now consider the case where ν, μ are finite measures. We begin with a preliminary lemma.

Lemma 4.9 *If μ and ν are finite measures and are not mutually singular then there is a measurable set A and a positive constant ϵ such that $\mu(A) > 0$ and $\forall E \in \Sigma.(E \subseteq A) \implies \epsilon\mu(E) \leq \nu(E)$.*

Proof. (Of lemma) Consider the set function $\nu - \frac{1}{n}\mu$ where n is a positive integer. Let P_n, N_n be the Hahn decomposition of this set function into its positive and negative sets respectively. If we can show that we have some P_n with $\mu(P_n) > 0$ we are done; we just have to choose $\epsilon = \frac{1}{n}$ and $A = P_n$. Now consider the set $M = \cup_n P_n$; if $\mu(M) > 0$ we are done since this can only happen if at least one of the P_n has positive μ-measure. Now $M^c = \cap_n N_n$ so M^c is a negative set for all the set functions $nu - \frac{1}{n}\mu$; i.e. we have $\forall n.\nu(M^c) \leq \frac{1}{n}\mu(M^c)$ which is only possible if $\nu(M^c) = 0$. Thus if $\mu(M) = 0$ we would have that μ and ν are mutually singular. Hence $\mu(M) > 0$. □

We continue with the proof of the main theorem. The strategy is to try to construct the f of part (2) by a limiting argument. We define $\mathcal{G}$ to be the set of functions g such that

$$\forall E \in \Sigma. \int_E g\mu \leq \nu(E).$$

This set has two important closure properties. First, if g_1, g_2 are both in $\mathcal{G}$ then $max(g_1, g_2)$ is also in $\mathcal{G}$. Second, if $g_n \uparrow g$ and all the g_n are in $\mathcal{G}$ then so is g. Both of these are easy exercises.

Now let $K = sup\{\int g\mu : g \in \mathcal{G}\}$ which exists since the set is bounded by $\nu(X)$; in particular $K \leq \nu(X)$. We are going to try to construct an f which attains this limiting value K when integrated over X. The fact that K may be less than $\nu(X)$ will account for the part of ν that is mutually singular with respect to μ. Let g_n be functions in $\mathcal{G}$ such that $\int_X g_n\mu > (K - \frac{1}{n})$; we know that we can find such functions because of the definition of K as a sup over $\mathcal{G}$. Now we let $f_n = max\{g_1, \ldots, g_n\}$; by the first closure property these functions are in $\mathcal{G}$. The sequence f_n is clearly increasing so if we define $f = \lim_{n \longrightarrow \infty} f_n$ we have $f_n \uparrow f$. Hence by the second closure property of $\mathcal{G}$ we have $f \in \mathcal{G}$. Thus by the monotone convergence theorem we get

$$K \geq \int_X f\mu = \lim_{n \longrightarrow \infty} \int_X f_n\mu \geq \lim_{n \longrightarrow \infty} \int_X g_n\mu = K.$$

Thus we have $\int_X f\mu = K$.

Now we can define the Lebesgue decomposition of ν. We set $\nu_a(E) = \int_E f\mu$ for any $E \in \Sigma$ and $\nu_s(E) = \nu(E) - \nu_a(E)$. It is easy to verify that ν_a and ν_s are measures and obviously $\nu_a << \mu$. We claim that $\nu_s \perp \mu$. Suppose not, then from the lemma we can find a set A and a number ϵ with the property that for any measurable set $E \subseteq A$ we have $\epsilon\mu(E) \leq \nu_s(E)$ as described in the lemma. We choose E to be any measurable set contained in A and get

$$\int_E (f+\epsilon\chi_A)\mu = \int_E f\mu+\epsilon\int_E \chi_A\mu = \int_E f\mu+\epsilon\mu(E) \leq \int_E f\mu+\nu_s(E) = \nu(E).$$

This means that $f' = f + \epsilon\chi_A$ is also in $\mathcal{G}$. The lemma says that $\mu(A) > 0$ thus $\int_X f'\mu = K + \epsilon\mu(A) > K$, but this contradicts the definition of K. Thus $\nu_s \perp \mu$.

We have both parts of the theorem for the case of finite measures and part (2) for σ-finite measures. To see that part (1) also extends to the σ-finite case we use part (2). We define $\lambda = \mu + \nu$ then clearly $\mu, \nu << \lambda$ so we have functions f, g such that $\int_E f\lambda = \mu(E)$ and $\int_E g\lambda = \nu(E)$. Let $B = \{x : f(x) > 0\}$ and set $\nu_a(E) = \nu(E \cap B)$ and $\nu_s(E) = \nu(E \cap B^c)$. Clearly we have $\nu = \nu_a + \nu_s$. Suppose that $\mu(A) = 0$ then $\int_A f\lambda = 0$, hence $\int_{A\cap B} f\lambda = 0$. But $f > 0$ on $A \cap B$, so $\lambda(A \cap B) = 0$, hence $\nu_a(A) = 0$. This shows that $\nu_a << \mu$. Now we note that $\nu_s(B) = \nu_s(B \cap B^c) = 0$ and $\mu(B^c) = \int_{B^c} f\lambda = 0$, thus $\nu_s \perp \mu$. □

4.4 Exercises

(1) Verify the two closure properties of $\mathcal{G}$ used in the proof of the Radon-Nikodym theorem.

(2) Give an example showing that the Radon-Nikodym theorem does not hold if the measures are not assumed σ-finite. [Hint: Use counting measure and Lebesgue measure on the real line.]

(3) We used the Hahn decomposition to prove one of the lemmas needed in the proof of the Radon-Nikodym theorem. Suppose that we have proved the Radon-Nikodym theorem some other way, derive the Hahn decomposition as a consequence.

Chapter 5

A Category of Stochastic Relations

In this chapter we discuss some categorical constructions which allow us to unify some of the ideas in probabilistic semantics. Readers interested in probabilistic process algebra or approximation theory need not read this chapter.

The fundamental idea, to look for a monad which imitates some of the properties of the powerset monad, goes back to Lawvere [Law63] but the detailed development is due to Giry [Gir81]. We have, however, modified her definition slightly and, in doing so, produced an example of a partially-additive category [MA86]. This connection allows for a simple presentation of Kozen's probabilistic semantics for a language of while loops [Koz81; Koz85]. The material in this chapter is mostly taken from [Pan99]. The name **SRel** is an evocation of the analogy between relations and Markov kernels.

5.1 The Category SRel

Definition 5.1 The precategory **SRel** has as objects (X, Σ_X) sets equipped with a σ-field. The morphisms are conditional probability densities or stochastic kernels. More precisely, a morphism from (X, Σ_X) to (Y, Σ_Y) is a function $h : X \times \Sigma_Y \longrightarrow [0, 1]$ such that

(1) $\forall B \in \Sigma_Y . \lambda x \in X . h(x, B)$ is a bounded measurable function,
(2) $\forall x \in X . \lambda B \in \Sigma_Y . h(x, B)$ is a subprobability measure on Σ_Y.

The composition rule is as follows. Suppose that h is as above and $k : (Y, \Sigma_Y) \longrightarrow (Z, \Sigma_Z)$. Then we define $k \circ h : (X, \Sigma_X) \longrightarrow (Z, \Sigma_Z)$ by the formula $(k \circ h)(x, C) = \int_Y k(y, C) h(x, dy)$.

It is clear that the composition formula really defines a morphism with the required properties; if it is not clear the reader should check it at this point.

This is very close to Giry's definition except that we have a subprobability measure rather than a probability measure. Henceforth, we write simply X for an object in **SRel** rather than (X, Σ_X) unless we really need to emphasise the σ-field. Before proceeding we prove

Proposition 5.2 *With composition defined as above* **SRel** *is a category.*

Proof. We use h, k to stand for generic morphisms of type X to Y and Y to Z respectively. We write A, B, C for measurable subsets of X, Y, Z respectively. The identity morphism on X is the Dirac delta "function", $\delta(x, A)$. The fact that it is the identity is simply the equation

$$h(x, B) = \int_X h(x', B)\delta(x, dx')$$

which we have verified before as our very first computation of a Lebesgue integral.

To verify associativity is a routine use of the monotone convergence theorem. Suppose h, k are as above and that $p : Z \longrightarrow W$ is a morphism and D is a measurable subset of W, we have to show

$$\int_Y [\int_Z p(z, D)k(y, dz)]h(x, dy) = \int_Z p(z, D)[\int_Y k(y, dz)h(x, dy)].$$

The free variables in the above are x and D. Note that this is not just a Fubini type rearrangement of order of integration, the role of the stochastic kernels change. On the lhs the expression in square brackets produces a measurable function of z, for a fixed D, this measurable function is the integrand for the outer (Y) integration and the measure for this integration is $h(x, dy)$. On the rhs the expression in square brackets defines a measure on Σ_Z which is used to integrate the measurable function $p(z, D)$ over Z. Now the above equation is just a special instance of the equation

$$\int_Y [\int_Z P(z)k(y, dz)]h(x, dy) = \int_Z P(z)[\int_Y k(y, dz)h(x, dy)]$$

where $P(z)$ is an arbitrary real-valued measurable function on Z. To prove this equation we need only verify it for the very special case of a characteristic function χ_C for some measurable subset C of Z. With $P = \chi_C$ we argue as follows. Recall that whenever we integrate a characteristic function χ_C wrt any measure ν we get $\nu(C)$. Thus on the lhs the expression in square brackets becomes $k(y, C)$ and the overall expression is

$\int_Y k(y,C)h(x,dy)$. On the rhs the result is the measure evaluated on C, in other words the expression in square brackets evaluated at C. This is exactly $\int_Y k(y,C)h(x,dy)$. The proof is now routinely completed by first invoking linearity to conclude that the required equation holds for any simple function and then the monotone convergence theorem to conclude that it holds for any measurable function. □

5.2 Probability Monads

In what sense are we entitled to think of the category **SRel** as a category of relations? It has a peculiarly asymmetric character and lacks some of the key properties associated with a category of relations, in particular closed structure, as we will discuss in the next section. There is, however, one way in which it does resemble the category of relations. Recall that the category of relations is the Kleisli category of the powerset functor over the category of sets. It turns out that **SRel** is the Kleisli category of a functor, which resembles the powerset functor, over the category **Mes** of measurable spaces and measurable functions.

We define the functor $\Pi : \mathbf{Mes} \longrightarrow \mathbf{Mes}$ as follows. On objects

$$\Pi(X) =_{df} \{\nu | \nu \text{ is a subprobability measure on} X\}.$$

For any $A \in \Sigma_X$ we get a function $p_A : \Pi(X) \longrightarrow [0,1]$ given by $p_A(\nu) =_{df} \nu(A)$. The σ-field structure on $\Pi(X)$ is the least σ-field such that all the p_A maps are measurable. A measurable function $f : X \longrightarrow Y$ becomes $\Pi(f)(\nu) = \nu \circ f^{-1}$. Checking that Π is a functor is trivial. Note the sense in which one can think of $\Pi(X)$ as the collection of probabilistic subsets (or "fuzzy" subsets) of X.

We claim that Π is a monad. We define the appropriate natural transformations $\eta : I \longrightarrow \Pi$ and $\mu : \Pi^2 \longrightarrow \Pi$[1] as follows:

$$\eta_X(x) = \delta(x,\cdot), \mu_X(\Omega) = \lambda B \in \Sigma_X. \int_{\Pi(X)} p_B \Omega.$$

The definition of η should be clear but the definition of μ needs to be deconstructed. First note that Ω is a measure on $\Pi(X)$. Recall that p_B is the measurable function, defined on $\Pi(X)$, which maps a measure ν to $\nu(B)$. The σ-field on $\Pi(X)$ has been defined precisely to make this a measurable function. Now the integral $\int_{\Pi(X)} p_B\Omega$ should be meaningful.

[1] Try not to confuse μ with a measure.

Of course one has to verify that $\mu_X(\Omega)$ is a subprobability measure. The only subtlety is verifying that countable additivity holds which we leave as an exercise.

Theorem 5.3 *[Gìry] The triple* (Π, η, μ) *is a monad on* **Mes**.

Proof. We omit the verification that η_X and μ_X are morphisms. We begin by stating four facts we need in the proof. Let X and Y be measurable spaces and let x, y denote elements of X and Y respectively. Let $f : X \longrightarrow Y$ be measurable, $\nu \in \Pi(X)$, $\Omega \in \Pi^2(X)$ and P, Q be bounded real-valued measurable functions on X and Y respectively.

(1) $\int_Y Q\Pi(f)(\nu) = \int_X (Q \circ f)\nu$.
(2) $\int_X P\eta_X(x) = P(x)$.
(3) Given any real-valued measurable function P we define $\xi_P : \Pi(X) \longrightarrow [0,1]$ by $\forall \nu \in \Pi(X).\xi_P(\nu) = \int_X P\nu$. We claim that ξ_P is measurable.
(4) With ξ_P as above we have

$$\int_X P\mu_X(\Omega) = \int_{\Pi(X)} \xi_P \Omega.$$

The first item was our very first example application of the monotone convergence theorem. The second item is an immediate consequence of the properties of the Dirac delta function. We leave the third item as an exercise and verify the fourth.

First note that when P is χ_B then ξ_P is just p_B. Let P be χ_B for some measurable subset B of X. Now we have

$$\int_X P\mu_X(\Omega) = \int_B \mu_X(\Omega) = \mu_X(\Omega)(B) = \int_{\Pi(X)} p_B\Omega == \int_{\Pi(X)} \xi_P\Omega.$$

Thus we have the result for a characteristic function. By linearity it holds for any simple function. Now assume that there is a family of simple functions $s_i \uparrow P$. We have, by the monotone convergence theorem

$$\int_X P\mu_X(\Omega) = \lim_{i \longrightarrow \infty} \int_X s_i\mu_X(\Omega).$$

But we know that this is equal to

$$\lim_{i \longrightarrow \infty} \int_{\Pi(X)} \xi_i\Omega$$

where ξ_i means xi_{s_i}. Now it is easy to see that $\lim_{i \longrightarrow \infty} \xi_i = \xi_P$[2] so by the monotone convergence theorem we get the result we want.

Now to prove that we have a monad we need to check the naturality of η and μ. The naturality of η is trivial from fact 2. The naturality of μ follows from an easy calculation with fact 1 used at the evident place. The verification of the triangle identity is a routine exercise as it just uses the definitions and no subtleties arise. We check the associativity equation explicitly. Let $\Omega' \in \Pi^3(X)$ and $B \in \Sigma_X$. We calculate

$(\mu_X \circ \Pi(\mu_X))(\Omega')(B)$

$\qquad = (\mu_X(\Pi(\mu_X)(\Omega')))(B)$

from the definition of μ_X we get

$\qquad = \int_{\Pi(X)} p_B \Pi(\mu_X)(\Omega')$

using fact 1 we get

$\qquad = \int_{\Pi^2(X)} p_B \circ \mu_X \Omega'$

from the definition of ξ we get

$\qquad = \int_{\Pi^2(X)} \xi_{p_B} \Omega'.$

In the other direction we calculate as follows

$(\mu_X \circ \mu_{\Pi(X)})(\Omega')(B)$

$\qquad = \mu_X(\mu_{\Pi(X)}(\Omega'))(B)$

from the definition of μ_X

$\qquad = \int_{\Pi(X)} p_B \mu_{\Pi(X)}(\Omega')$

using fact 4 we get

$\qquad = \int_{\Pi^2(X)} \xi_{p_B} \Omega'$

which is exactly what we got before. □

Now that we have that Π is a monad we can investigate the Kleisli category. A map, $X \longrightarrow Y$, in this category would be a map $X \longrightarrow \Pi Y$ in **Mes**. But if we recall that ΠY is $\Sigma_Y \longrightarrow [0,1]$ then by uncurrying we can write a Kleisli map as $X \times \Sigma_Y \longrightarrow [0,1]$, i.e. precisely the type of the morphisms in **SRel**. Of course one has to verify that one gets exactly the **SRel** morphisms. We leave this as an exercise.

[2] Check it if you don't believe it!

5.3 The Structure of SRel

We will examine the properties of the category **SRel**, especially the **partially additive** structure [MA86].

We begin by establishing that **SRel** has countable coproducts.

Proposition 5.4 *The category* **SRel** *has countable coproducts.*

Proof. Given a countable family $\{(X_i, \Sigma_I) | i \in I\}$ of objects in **SRel** we define (X, Σ) as follows. As a set X is just the disjoint union of the X_i. We write the pair (x, i) for an element of X, where the second member of the pair is a "tag", i.e. an element of I, which indicates which summand the element x is drawn from. The σ-field on X is generated by the measurable sets of each summand. Thus, a generic measurable set in X will be of the form $\uplus_{i\in I} A_i \times \{i\}$, where each A_i is in Σ_i. We will usually just write $\uplus_{i\in I} A_i$ with the manipulation of tags ignored when we are talking about measurable sets.

This object will be "the" coproduct in **SRel**. The injections $\iota_i : X_i \longrightarrow X$ are $\iota_i(x, \uplus_{k\in I} A_k) = \delta((x, i), \uplus_{k\in I} A_k) = \delta(x, A_i)$. Given a family $f_j : X_j \longrightarrow (Y, \Sigma_Y)$ of **SRel** morphisms we construct the mediating morphism $f : X \longrightarrow Y$ by $f((x, i), B) = f_i(x, B)$. We check the required commutativity by calculating

$$(f \circ \iota_j)(x_j, B) = \int_X f(x, B)\delta((x_j, j), \cdot) = \int_{X_j} f_j(x, B)\delta(x_j, \cdot) = f_j(x_j, B).$$

This is clearly the only way to construct f and satisfy all the required commutativities. □

This is very analogous to the construction in **Rel** but there the coproduct is actually a biproduct (since **Rel** is a self-dual category). This coproduct is not a biproduct. In fact it has a kind of restricted universality property that we will explain after we have discussed the partially additive structure of **SRel**.

It is easy to define a symmetric tensor product. Given (X, Σ_X) and (Y, Σ_Y) we define $(X, \Sigma_X) \otimes (Y, \Sigma_Y)$ as $(X \times Y, \Sigma_X \otimes \Sigma_Y)$ where we mean the tensor product of σ-fields defined earlier and cartesian product of the sets of course. We write $X \otimes Y$ to be brief. Given $f : X \longrightarrow X'$ and $g : Y \longrightarrow Y'$ we define $f \otimes g : X \otimes Y \longrightarrow X' \otimes Y'$ by

$$(f \otimes g)((x, y), A' \times B') = f(x, A')g(y, B')$$

where A' and B' are measurable subsets of X' and Y' respectively. Of course this defines it only on rectangles, but the collections of rectangles is a semi-ring and we can extend the measure to all measurable subsets of $X' \times Y'$. It is easy to see that one can define a symmetry.

In **Rel** we actually have a compact closed category in which the internal hom and the tensor coincide, which is a very special situation. In **SRel**, though, the tensor is exactly the same as in **Rel**, we do not even get closed structure. The reader should try to construct what seems at first sight to be the evident evaluation and coevaluation maps and see what fails. Roughly speaking one gets stuck at the point where one is required to manufacture a canonical measure on a σ-field; the only obvious candidate, the counting measure miserably fails to satisfy the required equations.

5.3.1 *Partially additive structure*

This subsection is a summary of the definitions of partially additive structure due to Manes and Arbib [MA86]. Their exposition concentrates on examples like partial functions. The category **SRel** provides a very nice example of their theory. Given $f, g : X \longrightarrow Y$ in **SRel** we can *sometimes* add them by writing $(f+g)(x,B) = f(x,B) + g(x,B)$. It may happen that the sum exceeds 1 in which case it is not defined, but if the sum $f(x,Y) + g(x,Y)$ is bounded by 1 for all x then we get a well-defined subprobability measure and a natural notion of adding morphisms. This is exactly the type of situation axiomatised in the theory of partially additive categories.

Definition 5.5 A **partially additive monoid** is a pair $(M, \sum)$ where M is a nonempty set and $\sum$ is a partial function which maps *some* countable subsets of M to M. We say that $\{x_i | i \in I\}$ is **summable** if $\sum_{i \in I} x_i$ is defined. The following axioms are obeyed.

(1) **Partition-Associativity:** Suppose that $\{x_i | i \in I\}$ is a countable family and $\{I_j | j \in J\}$ is a countable partition of I. Then $\{x_i | i \in I\}$ is summable iff for every $j \in J$ $\{x_i | i \in I_j\}$ is summable and $\{\sum_{i \in I_j} x_i | j \in J\}$ is summable. In this case we require

$$\sum_{i \in I} x_i = \sum_{j \in J} \sum_{i \in I_j} x_i.$$

(2) **Unary-sum:** A singleton family is always summable.

(3) **Limit:** If $\{x_i | i \in I\}$ is countable and *every finite subfamily* is summable then the whole family is summable.

One can think of this as axiomatising an abstract notion of convergence. However the first axiom says, in effect, that we are working with *absolute* convergence and hence rearrangements of any kind are permitted once we know that a sum is defined. Note that one can have some finite sums undefined and some infinite sums defined. The usual notion of complete partial order with sup as sum gives a model of these axioms. A vector space gives a typical nonexample as the limit axiom fails.

We state a simple proposition without proof.

Proposition 5.6 *The sum of the empty family exists, call it* 0*. It is the identity for* $\sum$*.*

Though this proposition is easy to prove it has important consequences as we shall see presently.

Definition 5.7 Let $\mathcal{C}$ be a category. A **partially additive structure** on $\mathcal{C}$ is a partially additive monoid structure on the homsets of $\mathcal{C}$ such that if $\{f_i : X \longrightarrow Y | i \in I\}$ is summable, then $\forall W, Z, g : W \longrightarrow X, h : Y \longrightarrow Z$, we have that $\{h \circ f_i | i \in I\}$ and $\{f_i \circ g | i \in I\}$ are summable and, furthermore, the equations

$$h \circ \sum_{i \in I} f_i = \sum_{i \in I} h \circ f_i, (\sum_{i \in I} f_i) \circ g = \sum_{i \in I} f_i \circ g$$

hold.

Since any partially additive monoid has a zero element, a category with partially additive structure will have “zero” morphisms.

Definition 5.8 A category has **zero morphisms** if there is a distinguished morphism in every homset – we write 0_{XY} for the distinguished member of $hom(X, Y)$ – such that $\forall W, X, Y, Z, f : W \longrightarrow X, g : Z \longrightarrow Y$ we have $g \circ 0_{WZ} = 0_{XY} \circ f$.

Proposition 5.9 *If a category has a partially additive structure it has zero morphisms.*

This follows immediately from proposition 5.6. Note that if a category has a partially additive structure then every homset is nonempty. This immediately rules out, for example, **Set** as a category that could support a partially additive structure.

Proposition 5.10 *The category* **SRel** *has a partially additive structure.*

Proof. A family $\{h_i : X \longrightarrow Y | i \in I\}$ in **SRel** is summable if

$$\forall x \in X. \sum h_i(x, Y) \leq 1.$$

We define the sum by the evident pointwise formula. Partition associativity follows immediately from the fact that we are dealing with absolute convergence since all the values are nonnegative. The unary sum axiom is immediate. To see the validity of the limit axiom we proceed as follows. Suppose that $\{h_i : X \longrightarrow Y | i \in I\}$ in **SRel** is summable, i.e. we assume that

$$\forall x \in X. \sum h_i(x, Y) \leq 1.$$

We define the sum by the evident pointwise formula. Partition associativity follows immediately from the fact that we are dealing with absolute convergence since all the values are nonnegative. The unary sum axiom is immediate. To see the validity of the limit axiom we proceed as follows. Suppose that $\{h_i : X \longrightarrow Y | i \in \mathbf{N}\}$ is a countable family and that every finite subfamily is summable. The sums $\sum_{i=1}^{n} h_i(x, Y)$ are bounded by 1 for all x. The sum $\sum_{i=1}^{\infty} h_i(x, Y)$ has to converge, being the limit of a bounded monotone sequence of reals and the sum has to be also bounded by 1. Thus the entire family is summable. One has to check that the sum of morphisms defined this way really gives a measure but the verification of countable additivity is easily done by using the fact that each h_i is countably additive and the sums in question can be rearranged since we have only nonnegative terms. The verification of the two distributivity equations is a, by now routine, use of the monotone convergence theorem mantra. □

We now define some morphisms which are of great importance in the theory of partially additive categories. They exist as soon as one has coproducts and a family of zero morphisms, thus they always exist in a category with partially additive structure.

Definition 5.11 Let $\mathcal{C}$ be a category with countable coproducts and zero morphisms and let $\{X_i | i \in I\}$ be a countable family of objects of $\mathcal{C}$.

(1) For any $J \subset I$ we define the **quasi-projection** $PR_J : \coprod_{i \in I} X_i \longrightarrow$

$\coprod_{j\in J} X_j$ by

$$PR_J \circ \iota_i = \begin{cases} \iota_i & i \in J \\ 0 & i \notin J \end{cases}$$

(2) We write $I \cdot X$ for the coproduct of $|I|$ copies of X. We define the **diagonal-injection** Δ by couniversality:

$$\begin{tikzcd} \coprod(X_i|i\in I) \arrow[r, "\Delta"] & I \cdot \coprod(X_i|i\in I) \\ X_j \arrow[u, "in_j"] \arrow[r, "in_j"'] & \coprod(X_i|i\in I) \arrow[u, "in_j"] \end{tikzcd}$$

(3) We have a morphism σ from $I \cdot X$ to X given by:

$$\begin{tikzcd} I \cdot X \arrow[r, "\sigma"] & X \\ X \arrow[u, "in_j"] \arrow[ur, "id_X"'] & \end{tikzcd}$$

These are all very simple maps to describe explicitly. In **Set** we cannot have a map which behaves like PR_J because we do not have zero morphisms. In **SRel** we have

$$PR_J((x,k), \uplus_{j\in J}) = \begin{cases} \delta(x, A_k) & k \in J \\ 0 & k \notin J \end{cases}.$$

The Δ morphism in **SRel** is

$$\Delta((x,k), \uplus_{i\in I}(\uplus_{j\in I} A^i_j)) = \delta(x, A^k_k).$$

The analogous map in **Set** is $\Delta((x,k)) = ((x,k),k)$.

Finally

$$\sigma((x,k),A) = \delta(x,A)$$

in **SRel** while in **Set** we have $\sigma((x,k)) = x$.

We are finally ready to define a partially additive category.

Definition 5.12 A **partially additive category**, $\mathcal{C}$, is a category with countable coproducts and a partially additive structure satisfying the following two axioms.

(1) **Compatible sum axiom**: If $\{f_i | i \in I\}$ is a countable set of morphisms in $\mathcal{C}(X, Y)$ and there is a morphism $f : X \longrightarrow I \cdot Y$ with $PR_i \circ f = f_i$ then $\{f_i | i \in I\}$ is summable.

(2) **Untying axiom**: If $f, g : X \longrightarrow Y$ are summable then $\iota_1 \circ f$ and $\iota_2 \circ g$ are summable as morphisms from X to $Y + Y$.

The first axiom says that if a family of morphisms can be "bundled together as a morphism into the copower" then the family is summable. The reverse direction is an easy consequence of the definition of partially additive structure so this is really an if and only if statement in a partially additive category.

Proposition 5.13 *The category* **SRel** *is a partially additive category.*

Proof. We already know that **SRel** has a partially additive structure and has countable coproducts. Suppose that we have the morphisms f_i and f as described in the compatible sum axiom. We verify that the f_i form a summable family. For fixed $x \in X$ and $B \in \Sigma_Y$ we have

$$\begin{aligned}\textstyle\sum_{i\in I} f_i(x, B) &= \textstyle\sum_{i\in I}(PR_i \circ f)(x, B)\\ &= \textstyle\sum_{i\in I} \int_{I \cdot Y} PR_i(u, B) f(x, du)\\ &= \textstyle\sum_{i\in I} \int_Y \chi_B(u) f(x, du)\end{aligned}$$

(in the previous line the integral is over the ith summand of the disjoint union only)

$$= \textstyle\sum_{i\in I} f(x, \iota_i(B)) = f(x, I \cdot B).$$

In the last line $I \cdot B$ means the disjoint union of $|I|$ many copies of B. From this calculation and the fact that f is a morphism in **SRel** we see that the sum is indeed defined. To verify untying is a very easy exercise. □

5.4 Kozen Semantics and Duality

In this short section we explain the point of the long digression into partially additive categories. Briefly, the point is to support a notion of iteration. We give a simple presentation of Kozen's probabilistic semantics for a language of while loops using the fact that **SRel** supports iteration simply by being a partially additive category. We first prove that there is an iteration operation whenever we have a partially additive category and then give the semantics. Kozen's first presentation was much more elaborate, but in a later paper he sketched essentially this semantics and described a very nice duality theory which gives a notion of probabilistic predicate transformers.

Theorem 5.14 *[**Arbib-Manes**] Given $f : X \longrightarrow X + Y$ in a partially additive category, we can find a unique $f_1 : X \longrightarrow X$ and $f_2 : X \longrightarrow Y$ such that $f = \iota_1 \circ f_1 + \iota_2 \circ f_2$. Furthermore there is a morphism $\dagger f =_{df} \sum_{n=0}^{\infty} f_2 \circ f_1^n : X \longrightarrow Y$. The morphism $\dagger f$ is called the **iterate** of f.*

Proof. The first assertion is trivial. We have $f_1 = PR_X \circ f$ and $f_2 = PR_Y \circ f$ where the PR maps are the ones associated with the coproduct $X + Y$. The second assertion is about the specific family $\{f_2 \circ f_1^n | n \geq 0\}$ being summable. We first prove by induction on k that the finite families $\{f_2 \circ f_1^n | k \geq n \geq 0\}$ are summable and the result then follows from the limit axiom. The base case is just the unary sum axiom applied to f_2. For the inductive step we claim that if $g : X \longrightarrow Y$ is any morphism then $g \circ f_1$ and f_2 are summable. The induction step then follows immediately from the claim by using $\sum_{n=0}^{k} f_2 \circ f_1^n$ for g. To prove the claim we note

$$\begin{aligned}[g, I_Y] \circ f &= [g, I_Y] \circ (\iota_1 \circ f_1 + \iota_2 \circ f_2) \\ &= [g, I_Y] \circ \iota_1 \circ f_1 + [g, I_Y] \circ \iota_2 \circ f_2 \\ &= g \circ f_1 + f_2\end{aligned}$$

Thus the claim is proved. □

More can be said about the iteration construct, in fact Bloom and Esik have written a monumental treatise on this topic and compared various axiomatisations of iteration. Iteration is closely linked to the notion of trace and is also the dual of a fixed-point combinator. We will not discuss the various equational properties of iteration except to note the fixed point property: given any $g : X \longrightarrow X$ we have $\dagger([g, I_Y] \circ f) = \dagger(f \circ g)$.

5.4.1 *While loops in a probabilistic framework*

We define the syntax as follows:

$$S ::== x_i := f(\vec{x}) | S_1; S_2 | if\ \mathbf{B}\ then\ S_1\ else\ S_2 | while\ \mathbf{B}\ do\ S.$$

We use the following conventions. We assume that the program has a fixed set of variables $\vec{x}$, say n distinct variables, and that they each take values in some measure space (X, Σ). The space (X^n, Σ^n) is the product space where the vector of variables takes its values. We assume that the function f is a measurable function of type $(X^n, \Sigma^n) \longrightarrow (X, \Sigma)$ and that $\mathbf{B}$ defines a measurable subset of (X^n, Σ^n). We can thus suppress syntactic details about expressions and boolean expressions. It is easy to extend what follows to cover variables of different sorts and to add random assignment.

We model statements in this programming language as **SRel** morphisms of type $(X^n, \Sigma^n) \longrightarrow (X^n, \Sigma^n)$. We write $\vec{A}$ for the product $A_1 \times \ldots \times A_n$.
Assignment: $x := f(\vec{x})$

$$\llbracket x_i := f(\vec{x}) \rrbracket(\vec{x}, \vec{A}) = \delta(x_1, A_1) \ldots \delta(x_{i-1}, A_{i-1}) \delta(f(\vec{x}), A_i) \\ \times \delta(x_{i+1}, A_{i+1}) \ldots \delta(x_n, A_n)$$

Sequential Composition: $S_1; S_2$

$$\llbracket S_1; S_2 \rrbracket = \llbracket S_2 \rrbracket \circ \llbracket S_1 \rrbracket$$

where the composition on the rhs is the composition in **SRel**.
Conditionals: *if* **B** *then* S_1 *else* S_2

$$\llbracket \textit{if } \mathbf{B} \textit{ then } S_1 \textit{ else } S_2 \rrbracket(\vec{x}, \vec{A}) = \delta(\vec{x}, \mathbf{B}) \llbracket S_1 \rrbracket(\vec{x}, \vec{A}) + \delta(\vec{x}, \mathbf{B}) \llbracket S_2 \rrbracket(\vec{x}, \vec{A})$$

where **B** denotes the complement of **B**.
While Loops: *while* **B** *do* S

$$\llbracket \textit{while } \mathbf{B} \textit{ do } S \rrbracket = h^{\dagger}$$

where we are using the † in **SRel** and the morphism $h : (X^n, \Sigma^n) \longrightarrow (X^n, \Sigma^n) + (X^n, \Sigma^n)$ is given by

$$h(\vec{x}, \vec{A_1} \uplus \vec{A_2}) = \delta(\vec{x}, \mathbf{B}) \llbracket S \rrbracket(\vec{x}, \vec{A_1}) + \delta(\vec{x}, \mathbf{B}^c) \delta(\vec{x}, \vec{A_2}).$$

The opposite category can be used as the basis for a "predicate transformer" semantics due to Kozen [Koz85]. We sketch the ideas briefly, a detailed exposition would require an excursion into Banach spaces and the topology of these spaces. This part is not self-contained but the reader can still get a good idea of how the construction works.

Definition 5.15 The category **SPT** has as objects sets equipped with a σ-field. Given a σ-field we obtain the Banach space of bounded, real-valued, measurable functions defined on X and denoted $\mathcal{F}(X)$. The sup defines the norm. A morphism $\alpha : X \longrightarrow Y$ in the category is a linear, continuous function $\alpha : \mathcal{F}(X) \longrightarrow \mathcal{F}(Y)$.

Theorem 5.16 ***[Kozen]***

$$\mathbf{SRel}^{op} \equiv \mathbf{SPT}.$$

Proof. (sketch) Given $h : X \longrightarrow Y$ in **SRel** we construct $\alpha_h : \mathcal{F}(Y) \longrightarrow \mathcal{F}(X)$ as follows:

$$\alpha_h = \lambda g \in \mathcal{F}(Y).\lambda x \in X. \int_Y g(y)h(x, dy).$$

One has to check that this is linear (clear) and continuous.

Given $\alpha : X \longrightarrow Y$ in SPT we construct $h : Y \longrightarrow X$ in **SRel** as follows:

$$h(y, A) = \alpha(\chi_A)(y).$$

We check that these maps are really inverses. Suppose that we start with an **SRel** morphism $h : X \longrightarrow Y$ and we construct α_h and then go back to **SRel** obtaining a stochastc kernel k. We have $k(x, B) = \alpha_h(\chi_B)(x)$ but by definition of α_h this is $\int_Y \chi_B(y)h(x, dy) = h(x, B)$. Thus we get back our original morphism. The other direction is not quite so trivial. Suppose that we start with an α, construct an h and then α_h. We have to show that for any $f \in \mathcal{F}(X)$ that $\alpha(f) = \alpha_h(f)$. Now we take the special case of a characteristic function χ_A for f. We have then $\alpha_h(\chi_A)(y) = \int_X \chi_A h(y, dx) = h(y, A) = \alpha(\chi_A)(y)$. Thus the required equality holds for characteristic functions. Now we invoke the monotone convergence theorem mantra and see that it works for any measurable function. □

In the dual view being adopted here, a bounded, measurable function is the analogue of a predicate on the set of states. An **SRel** morphism is a state transformer while an **SPT** morphism is a predicate transformer. The role of a state is played by a measure on the set of traditional states. The satisfaction relation of ordinary predicates and states is replaced by the integral. Thus the measurable function (predicate) f (ϕ) is “satisfied” by the measure (state) μ (s) written $\int f\mu$ ($s \models \phi$) giving a value in $[0, 1]$ ($\{0, 1\}$).

5.5 Exercises

(1) Verify that the composition in **SRel** really does yield an **SRel** morphism.
(2) Verify that the expression for $\mu_X(\Omega)$ in Section 2 really does give a measure.
(3) Show that fact 3 at the start of the proof of Theorem 5.3 holds.
(4) Verify the triangle identity needed in the proof that Π is a monad.
(5) Exhibit an explicit adjunction between the categories **SRel** and **Mes**.

Chapter 6

Probability Theory on Continuous Spaces

In this chapter we use the mathematical tools that we have developed to formalise the basic ideas of probability theory in a way suitable for use in situations with continuous state spaces. For example, the concept of conditional probability density will be formalised using the Radon-Nikodym theorem. Good sources for this material are the books by Dudley [Dud89] or Billingsley [Bil95].

6.1 Probability Spaces

The basic arena for the study of probability is a *probability space*, this time equipped with a σ-algebra. We give the formal definition

Definition 6.1 A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set called the **sample space**, $\mathcal{F}$ is a σ-field on Ω and P is a probability measure on $\mathcal{F}$.

In the discrete case, where Ω is finite or countable, we took $\mathcal{F}$ to be the powerset of Ω and hence everything was measurable.

The intended meaning of a probability space is that one has a (one-step) process operating which ends up in a state or one is carrying out an experiment with the outcomes governed by some probabilistic process. The set Ω is the set of possible states or possible results. A member of $\mathcal{F}$ is called an *event*; not all subsets can be events anymore. The idea is that we cannot always tell, with our limited observational powers, exactly which point in Ω occurs; at best we may only be able to identify or specify some larger measurable set. We speak of an event σ *occurring* if the result is in the set σ. In continuous situations it typically happens that the singletons are measurable sets but that P ascribes 0 probability to them. Then we need to work with other measurable sets to say something quantitative.

6.2 Random Variables

Random variables are the actors that play in the theatre of the probability space. Mathematically they are just measurable functions but analysts and probabilists use different notations. Probability theory, however, developed independently of mathematical analysis for a long time – until Kolmogorov – and developed its own compelling metaphor and concomitant terminology.

Associated with the process described by a probability space are some measurable quantities – the random variables. For example, the probability space may be the state space of a chemical mixture and associated with it are some measurable physical quantities such as temperature and pressure – these are typical random variables. In most textbooks random variables are defined to take values in the real numbers. Conceptually, the theory is affected very little by defining a random variable to take values in any measure space, but important quantities, such as the expected values of random variables, which rely on the arithmetic of the reals, may not make sense. In these notes we will use the conventional textbook definition but occasionally more general random variables will arise.

Definition 6.2 A **random variable** on a probability space $(\Omega, \mathcal{F}, P)$ is a real-valued, *Borel measurable* function defined on Ω. A random variable which takes on values in the extended reals is called an **extended random variable**.

It is conventional to use uppercase letters like X for random variables rather than letters like f to suggest their role as functions.

The most important fact about random variables is that all the probabilistic information is captured in one function from $\mathbf{R}$ to $\mathbf{R}$ called the *distribution function.* Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$, fixed for the rest of the paragraph. Now given X we can define a probability measure, P_X on $\mathbf{R}^1$, by the formula

$$P_X(A) = P(\{\omega : X(\omega) \in A\}),$$

where A is a Borel set. Knowing this measure gives us all the information about the random variable. Now we can define a function $F_X : \mathbf{R} \longrightarrow [0, 1]$ by $F_X(x) = P_X$ by $F_X(x) = P(\{\omega : X(\omega) \leq x\}$, which captures

[1] We mean on the Borel sets of the reals.

all the information in the measure P_X. This function is increasing, right-continuous and satisfies

$$\lim_{x \longrightarrow \infty} F_X(x) = 1 \text{ and } \lim_{x \longrightarrow -\infty} F_X(x) = 0.$$

The notions of random variables and of distribution functions generalise in the obvious way to $\mathbf{R}^n$, but of course the computations are more intricate.

In chapter 1 we mentioned the familiar notion of independence of events.

Definition 6.3 We say a finite set $\{X_1, \dots, X_n\}$ of random variables defined on $(\Omega, \mathcal{F}, P)$ are **independent** if for all Borel sets $B_1, \dots, B_n$ we have

$$P(\{\omega : X_1(\omega) \in B_1, \dots, X_n(\omega) \in B_n\}) = \prod_{i=1}^{n} P(\{\omega : X_i(\omega) \in B_i\}).$$

This definition does not depend on the random variables taking values in the reals, thus it may be used for arbitrary measurable functions.

The most basic theorem about independence is the fact that the distribution function factorises.

Theorem 6.4 *Let $\{X_1 \dots, X_n\}$ be random variables on $(\Omega, \mathcal{F}, P)$ and let X be the ($\mathbf{R}^n$-valued) random variable $(X_1, \dots, X_n)$. Let the distribution functions be F_i and F respectively. Then*

$$F(x_1, \dots, x_n) = F_1(x_1) \dots F_n(x_n).$$

A key concept associated with random variables is the *expectation value.*

Definition 6.5 Given a random variable X on a probability space $(\Omega, \mathcal{F}, P)$, we define the **expectation value** $E[X]$ by the integral

$$E[X] = \int_{\Omega} X dP.$$

In the discrete case one has the usual summation formula for expectation value.

By applying the Fubini-Tonelli theorem we obtain the basic result about expectations of independent random variables.

Proposition 6.6 *Let $X_1, \dots, X_n$ be independent random variables defined on $(\Omega, \mathcal{F}, P)$. Assume that all X_i are nonnegative then*

$$E[X_1 \cdot X_2 \cdot \dots \cdot X_n] = E[X_1] \cdot \dots \cdot E[X_n].$$

There are many quantitative results about expectation values. A good survey, from the point of view of algorithms, is to be found in the recent book, *Randomized Algorithms* by Motwani and Raghavan [MR95].

6.3 Conditional Probability

In the introduction we discussed the concept of conditional probability in the discrete setting where we were able to give a satisfactory account without the apparatus of measure theory. In the continuous case the full arsenal of techniques that we have developed so far is needed. In particular the Radon-Nikodym theorem plays a crucial role in even defining conditional probability density.

Recall that conditional probability was defined as follows: given two events A and B, we write $P(A|B)$ for the conditional probability of A given B and define it as

$$P(A|B) = P(A \cap B)/P(B),$$

when $P(B) \neq 0$. In the continuous case we often have situations where B is a single point and more generally a set with $P(B) = 0$ so we cannot just hope to sidestep the situations where $P(B) = 0$. We can, however, consider the following heuristic approach; the line of argument is taken from Feller [Fel71]. We consider the case where B is a single point, say $\{b\}$ and we suppose that we have a nested family of sets $B_1 \supseteq B_2 \supseteq \dots B_i \supseteq \dots$ with $\cap B_i = \{b\}$. Now we can define $P(A|\{b\})$ as the limit of $P(A|B_i)$.

The above "definition" gives some intuition for the notion of conditional probability in the continuous case but it is very far from being formal. However it does point out what is needed, viz. some way of defining a "derivative of a measure" – precisely what the Radon-Nikodym theorem provides. Thus we expect to work with conditional probability density rather than with conditional probability. The density should try to mimic the discrete notion of conditional probability as far as possible.

Conditional probability describes how to revise estimates of probability given definite information – but, what sort of definite information do we consider? A reasonable starting point is to imagine the sample space Ω to be a product $X \times Y$ equipped with some joint probability distribution P on $X \times Y$. Of course, we are really talking about a probability distribution on the product σ-field but we will suppress mention of the σ-fields except where necessary. Now we can imagine that we know one of the coordinates precisely, say X, and we are interested in defining the conditional probability $P(B|X = x)$ which means that we wish to know the probability that a sample point with a known X-coordinate has a Y-coordinate in B. We will write $P(x, B)$ for this conditional probability. Given a joint probability distribution P we acquire two natural probability measures P_X, P_Y on Σ_X, Σ_Y respectively given by $P_X(A) = P(A \times Y)$ and $P_Y(B) = P(X \times B)$. It is reasonable to require that

$$\forall A \in \Sigma_X . P(A \times B) = \int P(x, B) dP_X,$$

and the symmetrical relation between the dual conditional probability density and P_Y. This requires that, for fixed B, the function $P(x, B)$ should be a measurable function.

How do we know that such a function can be found?

Proposition 6.7 *Given the product space $X \times Y$ of the previous paragraph with a joint probability distribution P, there are, for each fixed subset $B \in \Sigma_Y$ and for each fixed subset $A \in \Sigma_X$, measurable functions $P(x, B)$ and $P(y, A)$, such that*

$$P(A \times B) = \int P(x, B) dP_X = \int P(y, A) dP_Y.$$

These functions are unique P_X-almost (or P_Y-almost) everywhere.

This is an immediate consequence of the Radon-Nikodym theorem and is left as an easy exercise.

More generally, we can define a conditional probability density as follows. Suppose that we have a probability space (X, Σ_X, P), a measurable space (Y, Σ_Y) and a measurable function $f : X \longrightarrow Y$. We ask whether we can define a conditional probability of the form $P(A|y)$ where A is a measurable subset of X and y is a point in Y; the intended interpretation is the probability of a sample point x being in A given that $f(x) = y$. Clearly the situation in the last paragraph is a special case of this. Once again we have as a consequence of the Radon-Nikodym theorem,

Proposition 6.8 *Let X, Y, P, f be as in the discussion above. Let P_f be the probability measure $P \circ f^{-1}$ on Y induced by f. Then, for each $A \in \Sigma_X$, there is a P_f-almost everywhere unique measurable function, written $P(A|y)$, such that*

$$\forall B \in \Sigma_Y . P(f^{-1}(B) \cap A) = \int_B P(A|y) dP_f.$$

One can also define conditional expectations in essentially the same way. Suppose we have a probability space (X, Σ, P) and we have two random variables f and g defined on X. We can ask for the expectation value of g given that we know that f is r. The phrase "f is r" requires some explanation. We imagine that we perform repeated trials and get values for the random variables f and g. We select only those trials for which the value of the random variable f is r and compute the expectation value for g. We can impose a very similar requirement on conditional expectation values and once again appeal to the Radon-Nikodym theorem to conclude that a conditional expectation exists.

Proposition 6.9 *Let f be an extended random variable on (X, Σ_X, P) and $g : (X\Sigma_X) \longrightarrow (Y, \Sigma_Y)$ be a measurable function. Assume that $\mathbb{E}(f)$ exists. There is a an extended random variable h defined on Y such that for each $B \in \Sigma_Y$*

$$\int_{g^{-1}(B)} f dP = \int_B h dP_g$$

where $P_g = P \circ g^{-1}$. Furthermore, h is unique P_g-almost everywhere. We write $\mathbb{E}(g|f)$ for h.

Recall that we are referring to the Borel sets when we talk about random variables as measurable functions. Note that if we choose the function g to be a characteristic function then we immediately get conditional probability as a special case of conditional expectation.

Though conditional probability has a more direct intuition, we will talk about conditional expectation since one can obtain conditional probability results as a special case. The form in which we have developed conditional probability and conditional expectation is not very useful for technical arguments and a more general, though less intuitive, form is preferred in the probability literature. In order to motivate it consider what conditional probability is; it is a rule for revising one's original probability estimates in the face of new information. Now this information may take the general form of reporting the outcome of some experiment. We have looked at the special case where this information comes to us in the form of *precise information* about some variable; of course the process may depend on other variables that we know nothing about. In general we may merely know that a variable being tested for lies in some set or range instead of getting a precise value for it. Thus, in general, imagine that we are trying to measure some quantity and all we can tell is which of several sets of possible values the result is in. Then we can tell which sets the value lies in for any members of the σ-field generated by the given sets. We take this then as the basic datum. An experiment defines a σ-field and we can tell which members of the σ-field the result lies in. It may happen that the σ-field includes the singletons but it may not. This leads to the notion of *conditioning with respect to a σ-field.*

The situation one is interested in is the following. There is a probability space (X, Σ, P) and a subfield Λ of Σ is given. This subfield describes the experiment used to condition the probabilities and expectations. The precise theorem is as follows.

Theorem 6.10 *Let (X, Σ, P) be a probability space and let Λ be a subfield of Σ. Let f be an extended random variable on (X, Σ). Then there is a function $\mathbb{E}(f|\Lambda)$, which is a Λ-measurable extended real-valued function (random variable) defined on X such that*

$$\forall C \in \Lambda. \int_C f\, dP = \int_C \mathbb{E}(f|\Lambda) dP.$$

The proof is left as an exercise. It is important to note that the theorem is saying something nontrivial; it is easy to think that $\mathbb{E}(f|\Lambda)$ exists, "just choose f for it". This will not work, however, because the conditional expectation is required to be Λ-measurable and f is not going to be Λ-measurable in general. Of course the proofs of all these propositions are easy observations following from the Radon-Nikodym theorem but they are nevertheless nontrivial.

In just the same way one can prove an analogous result for conditional probability density given a σ-field.

Theorem 6.11 *Let (X, Σ, P) be a probability space and let Λ be a subfield of Σ. For each A, a member of Σ, there is a measurable function, P-almost everywhere unique, $P[A|\Lambda] : (X, \Lambda) \longrightarrow (\mathbf{R}, \mathcal{B})$, called the conditional probability of B given Λ, such that*

$$\forall C \in \Lambda . P(A \cap C) = \int_C P[A|\Lambda] dP.$$

One can prove analogues of most of the standard theorems for conditional probability density and conditional expectation. Thus there is a monotone convergence theorem, for example, for conditional expectations. A modern book like Dudley's [Dud89], Ash's [Ash72] or Billingsley's [Bil95] has a collection of such results.

6.4 Regular Conditional Probability

A disturbing feature of conditional probability is the dependence of the conditional probability density on the set whose probability is being expressed. Thus, for example, consider the statement in Theorem 6.11. The object $P[A|\Lambda]$ is a measurable function with a dependence on A. How does $P[\cdot|\Lambda](x)$ behave viewed as a set function with x fixed? Naively one would expect it to be a measure. If it were, then we would obtain a very pleasing object, called variously a conditional probability distribution, a stochastic kernel[2] or a Markov kernel. Such an object can be viewed as a two-argument function, one argument is a point and the other is a set. Fixing the set we get a measurable function, and fixing the point we get a measure. The term *kernel* is supposed to be evocative of kernels of integral operators. In the next two chapters we study objects of this kind from a categorical perspective.

Before we can study these objects we need to make sure that they exist and without further assumptions they do not! It is worth understanding why. Recall that the uniqueness of the conditional probability is guaranteed only modulo a set of measure 0. Now for a fixed countable family of pairwise disjoint sets a failure of countable additivity only occurs on a set of measure 0. Unfortunately there are uncountably many such families so we could have a failure of countable additivity everywhere! It turns out that assumptions essentially related to metric structure are necessary. In fact the theory of stochastic processes really relies quite heavily on convergence of measures on metric spaces following Kolmogorov's fundamental work in this field [Par67]. In this section we prove a standard result which guarantees the existence of regular conditional probability distributions on Polish spaces. In fact regular conditional probability distributions exist in considerably more general situations, for example on analytic spaces. The most general result is due to Jan Pachl [Pac78]; for a thorough treatment of regular conditional probability densities see [HJ94b].

We follow the exposition in Ash [Ash72]. The first step is to show that regular conditional distribution functions exist.

Definition 6.12 Let f be a random variable on (X, Σ, P) and let Λ be a subfield of Σ. We say that a function $F : X \times \mathbf{R} \longrightarrow [0, 1]$ is a **regular conditional distribution function** for f given Λ iff the following two conditions are satisfied:

[2] I have only seen this terminology in Feller.

(1) For each x, $F(x,\cdot)$ is a distribution function, i.e. increasing, right-continuous and

$$\lim_{r\longrightarrow\infty} F(x,r)=1 \text{ and } \lim_{r\longrightarrow-\infty} F(x,r)=0,$$

(2) $\forall r\in\mathbf{R}F(x,r)=P[f^{-1}((-\infty,r])|\Lambda](x)$.

The next theorem uses in an essential way the fact that the reals form a complete separable metric space.

Theorem 6.13 *If f is a random variable on a probability space (X,Σ,P) and Λ is a subfield of Σ then there is always a regular conditional distribution function for f given Λ.*

The strategy of the proof is to glue together the different versions of the conditional probability distribution. Since the reals form a separable metric space it suffices to work with a countable collection, indexed by the rationals. Now one identifies all the places where this collection violates the requirements of regular conditional distribution function and shows that they have measure 0. With this one can easily construct a regular conditional distribution function. In the proof given we omit some steps that involve knowing properties of conditional expectation.

Proof. For each real number r we know that there is a function $F_r=P[f^{-1}((-\infty,r])|\Lambda](x)$ from Theorem 6.11. There are many different versions of such a function, any pair of which differ only on a set of measure 0. Let $q_1,q_2,\ldots$ be an enumeration of the rationals. For each rational q we choose a version of F_q; we write F_i rather than F_{q_i}.

First we identify those places where the versions of the F_i that we have chosen fail to form an increasing function. We define the sets $A_{ij}=\{x: F_j(x)<F_i(x)\}$ and $A=\cup\{A_{ij}:q_i<q_j\}$. Now since $q_i<q_j$, a standard monotonicity property of conditional expectations says that $F_i\leq F_j$ P-almost everywhere. This means that each A_{ij} has measure 0 and hence A does too.

The next property we examine is right-continuity. We define B_i to be the set of points where F_i fails to be right-continuous and we define $B=\cup B_i$. By applying a conditional version of the dominated convergence theorem we conclude that each B_i and hence B has measure 0. Similarly we can use conditional versions of the monotone convergence theorem to show that the set $C=\{x:\lim_{n\longrightarrow\infty}F_n(x)\neq 1\}$ has measure 0 and that $D=\{x:\lim n\longrightarrow-\infty F_n(x)\neq 0\}$ also has measure 0. We refer to the union of A,B,C and D as E.

Now we can define our required F by

$$F(x,r) = \begin{cases} \lim_{q \longrightarrow r^+} F_q(x) & \text{if } x \notin E \\ \text{any proper distribution function} & \text{if } x \in E \end{cases}$$

Now it is easy to verify the properties of a regular conditional distribution function for f given Λ. □

The fact that the reals have a countable dense subset is used in the following way. By countability we can establish that the set of versions we are looking at misbehaves only on a set of measure 0. By the fact that this set is dense we can define the regular conditional distribution function almost everywhere by a limiting construction. This argument can essentially be carried out on any topological space that is the space underlying a complete separable metric space.

Definition 6.14 Let $f(X, \Sigma_X) \longrightarrow (X, \Sigma_Y)$ be a measurable function and Λ a sub-σ-field of Σ_X. The function $h : X \times \Sigma_Y \longrightarrow [0,1]$ is called a **regular conditional probability** for f given Λ if

(1) $\forall x \in X.h(x,\cdot)$ is a probability measure,
(2) $\forall B \in \Sigma_Y.h(x,B) = P[f^{-1}(B)|\Lambda](x)$.

Using the techniques for proving the existence of regular conditional distribution function one can prove the theorem below. Before we state it we define Polish spaces.

Definition 6.15 A **Polish space** is the topological space underlying a complete, separable metric space.

Recall that separable means that there is a countable base for the topology. This really abstracts the properties that we used in the proof above. The following theorem is not a surprise now.

Theorem 6.16 *Let $f : (X, \Sigma_X, P) \longrightarrow (Y, \Sigma_Y)$ be a measurable function and Y be a Polish space with Σ_Y its σ-field of Borel sets. Let Λ be a sub-σ-field of Σ_X. Then a regular conditional probability for f given Λ exists.*

The theorem can be proved quite a bit more generally than this. For example if Y is *Borel isomorphic*[3] to a Borel subset of a Polish space the theorem goes through.

[3] This means that there is a measurable bijection whose inverse is also measurable.

6.5 Stochastic Processes and Markov Processes

Roughly speaking a *stochastic process* is a probabilistic dynamic system. The word "dynamic" is supposed to convey the idea that there is some sort of temporal evolution. The mathematical theory is, however, stated rather more generally.

Definition 6.17 A **stochastic process** is an indexed family of random variables $X_t : \Omega \longrightarrow \mathbf{R}$ where $(\Omega, \mathcal{B}, P)$ is a probability space and $t \in T$ is the indexing set.

One usually thinks of T as "time", so it can be viewed as an ordered subset of the reals. In principle, the probability space can vary too but, for simplicity, we shall assume a fixed probability space.

Given this view, we can define the joint distribution $P_{t_1 \ldots t_n}$ of the variables $X_{t_1}, \ldots, X_{t_n}$ as a measure on $\mathbf{R}^n$,

$$P_{t_1 \ldots t_n}(B) = P(\{x | (X_{t_1}(x), \ldots, X_{t_n}(x)) \in B\}).$$

This satisfies the obvious consistency requirement – called the *Kolmogorov consistency requirement* – below

$$P_{t_1 \ldots t_n t_{n+1}}(B \times \mathbf{R}) = P_{t_1 \ldots t_n}(B).$$

This says that the last variable can be integrated out to give the prior distribution. Note that we do not intend the "time" to be discrete here. The second Kolmogorov consistency requirement states that if the variables are permuted then the distributions are altered in the obvious way. The fundamental theorem of the subject says that any family of finite dimensional probability distributions satisfying these two conditions can be realised as a stochastic process.

One can think of the probability distribution at time t, i.e.

$$P_t(A) = P(\{x | X_t(x) \in A\})$$

as representing the state of a transition system. The passage from P_t to P_s with $t < s$ may be thought of as a "transition" (discrete) or "evolution" of a system. In general, stochastic processes allow one to consider the possibility of the steps depending on the entire past history of the processes. A very important restriction is to consider processes where the transition probabilities depend on the entire evolution so far. A very important special class of stochastic processes are *Markov processes.* These are processes in which the transitions depend only on the current state.

More precisely we proceed as follows. We write $P(A_{n+1}|x_1,\ldots,x_n)$ for the conditional probability that the system is in the set A_{n+1} given that at time t_1 it was at x_1, etc. Now in a Markov process we have

$$P(A_{n+1}|x_1,\ldots,x_n) = P(A_{n+1}|x_n).$$

In other words, only the latest time matters. This definition applies equally well to discrete and continuous-time systems. This is a restriction, but a large number of systems are indeed found to be Markovian. Furthermore many apparently non-Markovian processes can be redefined to be Markovian by changing the state space. Thus if the transitions depend on a bounded number of past states the state space can be redefined to make it a Markov process by making the states tuples of former states. The key feature of a Markov process is that one can think of the transitions as being governed by a transition matrix (discrete state space) or stochastic kernel (continuous state space).

Typical examples of Markov processes are probabilistic automata, branching processes, random walks, arrival processes and a multitude of others of practical importance. The literature is vast and of varied accessibility. The standard probability texts contain references to the literature on this topic. Among many excellent places to start are the books by Kingman and Taylor [KT66] and by Billingsley [Bil95].

Chapter 7

Bisimulation for Labelled Markov Processes

In this chapter we introduce a class of labelled transition systems – labelled Markov processes (LMP) – and define bisimulation for them. Labelled Markov processes are probabilistic labelled transition systems where the state space is not necessarily discrete. The state space could be the reals, for example. The labels represent actions that the process could do, or be subject to. These are just like the actions in a Markov decision process (MDP), the only difference is one of viewpoint.

In both MDPs and LMPs one views the actions as "external" and indeterminate. When an action is performed a probabilistic transition occurs; the final state is a result of the choice of action and the internal indeterminacy of the system modelled by the probabilistic transition[1]. In LMPs one is interested in guarantees about the behaviour of the system under all possible external choices of the actions. The idea is that the system functions in an environment which chooses the actions and we want to ensure that the system satisfies behavioural specifications regardless of what the environment may do. In this model – the *reactive* model – we do not model how the external actions are chosen: we do not, for example, assign probabilities to external actions. LMPs were developed in the context of verification and the environment was viewed as an "adversary".

By contrast, MDPs were developed in the context of optimisation. One is interested in optimising some aspect of the behaviour assuming that one can choose the actions. One often assigns a reward to the actions or states of the system and the main interest is maximising the expected or average reward by formulating an appropriate *policy* for choosing actions. Thus

[1] Sometimes there is a further level of indeterminacy where after an action one of several possible probabilistic transitions is chosen. We will assume that this does not happen.

though the models are similar there are quite different viewpoints.

For the moment these differences will not be important because we are interested in the mathematics of the models. In a later chapter we will look at applications of the theory to MDPs.

The main point of this work is to understand how stochastic continuous systems interact with discrete systems. Typically one will have a piece of software connected to some physical device. Telecommunication or process control systems are good examples. To analyse such so-called “hybrid systems” one needs ideas from process algebra (bisimulation) as well as continuous mathematics. The fundamental new ingredient offered by computer science (apart from the manifest idea of computability or effectiveness) to these subjects is *compositionality.* Thus while the theory of stochastic processes [CM65] is concerned with a detailed analysis of the time evolution of systems behaving according to probabilistic laws, very little is ever done to analyse the behaviour of coupled systems in a systematic way. Computer scientists have stressed compositionality as a way to attack the formidable intricacy of the systems they have dealt with. For example, in Hillston’s work, compositionality is the key contribution of her approach to performance modelling [Hil94].

The notion of bisimulation is central to the study of concurrent systems. While there are a bewildering variety of different equivalence relations between processes (two-way simulation, trace equivalence, failures equivalence and many more) bisimulation enjoys some fundamental mathematical properties – most notably its characterisation as a fixed-point – which makes it the most discussed process equivalence. Of course there are many different variants of bisimulation itself! We are not concerned with adjudicating between the rival claims of all these relations, but rather, we are concerned with showing how to extend these ideas to the world of continuous state spaces. As we shall see below, new mathematical techniques (from the point of view of extant work in process algebra) have to be incorporated to do this. Once the model and the new mathematical ideas have been assimilated, the whole gamut of process equivalences can be studied and argued about.

From an immediate practical point of view, bisimulation can be used to reason about probabilistic, continuous state-space systems (henceforth Markov processes) in the following simple way. One often “discretises” a continuous system by partitioning the state space into a few equivalence classes. Usually one has some intuition that the resulting discrete system “behaves like” the original continuous system. This can be made precise by our notion of bisimulation. It is also the case that some systems cannot be so discretised, later we will develop approximation techniques to deal with these cases.

7.1 Ordinary Bisimulation

The central concept in this chapter is *bisimulation*, a behavioural equivalence between processes. Bisimulation was invented by David Park [Par81] as a formalisation of Milner's notion of behavioural equivalence [Mil80] in the context of process algebra. Park's original slides on this are hard to find and the main source for this material is Milner's book on CCS just cited. A fascinating historical account of bisimulation is available from Sangiorgi's web page [San04]. There were precursors of this idea in logic, automata theory and the theory of Markov chains [KS60].

In the case of labelled transition systems one can explain the concept as follows. Suppose that $(S, \mathcal{A}, \rightarrow \subseteq S \times \mathcal{A} \times S)$ is a labelled transition system. We say that an equivalence relation R is *a bisimulation* relation if whenever sRt we have

$$\forall a \in \mathcal{A}\ s \xrightarrow{a} s' \Rightarrow \exists t'(t \xrightarrow{a} t' \text{ and } s'Rt')$$

and *vice versa.* Thus this is a relation preserved by the dynamics. The apparent circularity in the definition is in fact not a circularity at all. This is simply a property of R which may or may not hold for a given R. We say s and s' are *bisimilar* – and write $s \sim s'$ – if there is a bisimulation relation R with sRs'. Thus, $\sim$ is the largest bisimulation relation.

It can be defined as the greatest fixed point of a suitable functional on the lattice of relations. Let $\mathcal{R}$ be the lattice of binary relations ordered by inclusion. We define $\mathcal{F} : \mathcal{R} \longrightarrow \mathcal{R}$ as follows:

$$s\mathcal{F}(R)t \iff \forall a \in \mathcal{A}\ s \xrightarrow{a} s' \Rightarrow (\exists t'\ t \xrightarrow{a} t' \text{ and } s'Rt')$$

and *vice versa.* It is easy to check that this functional is monotone on the lattice of relations and therefore has a greatest fixed point. The greatest fixed point is precisely the relation $\sim$ above. We will use both these viewpoints in our probabilistic extension.

7.2 Probabilistic Bisimulation for Discrete Systems

In this section, we recapitulate the Larsen-Skou definition of probabilistic bisimulation [LS91]. The systems that they consider will be called *labelled Markov chains* here.

Definition 7.1 A **labelled Markov chain** is a quadruple $(S, \mathcal{A}, C_a, P_a)$, where S is a countable set of states, $\mathcal{A}$ is a set of actions, and for each $a \in \mathcal{A}$, we have a subset C_a of S, a function, P_a, called a **transition probability matrix**,

$$P_a : C_a \times S \longrightarrow [0, 1]$$

satisfying the normalisation condition

$$\forall a \in \mathcal{A}, s \in C_a . \Sigma_{s' \in S} P_a(s, s') = 1.$$

If we have the weaker property,

$$\forall a \in \mathcal{A}, s \in C_a . \Sigma_{s' \in S} P_a(s, s') \leq 1$$

we call the system a **partial** labelled Markov chain.

The sets C_a are the sets of states that *can* do an a-action. If we have partial labelled Markov chains then we can just dispense with the C_a sets. In what follows we suppress the action set, i.e. we assume a fixed action set given once and for all.

Definition 7.2 Let $T = (S, P_a)$ be a labelled Markov chain. Then a **probabilistic bisimulation** $\equiv_p$, is an equivalence on S such that, whenever $s \equiv_p t$, the following holds:

$$\forall a \in \mathcal{A}. \forall A \in S/\equiv_p,\ \Sigma_{s' \in A} P_a(s, s') = \Sigma_{s' \in A} P_a(t, s').$$

Two states s and t are said to be **probabilistically bisimilar** ($s \sim_{LS} t$) in case (s, t) is contained in some probabilistic bisimulation.

Intuitively, we can read this as saying that two states are bisimilar if we get the same probability when we add up the transition probabilities to all the states in an equivalence class of bisimilar states. The addition is crucial – the probabilities are not just another label. The subtlety in the definition is that one has to somehow know what states are probabilistically bisimilar in order to know what the equivalence classes are, which in turn one needs in order to compute the probabilities to match them appropriately. Note the similarity of this definition with the first definition of ordinary bisimulation.

The paper by Larsen and Skou does much more than just define bisimulation. They introduce the notion of testing a probabilistic process and associating probabilities with the possible outcomes. They then introduce a notion of testable properties. The link with probabilistic bisimulation is that two processes are probabilistically bisimilar precisely when they agree with the results of all tests. They also introduce a probabilistic modal logic and show that bisimulation holds precisely when two processes satisfy the same formulas. We will not study testing in this book but it has been analysed by Stoelinga and Vaandrager [SV03] and by van Breugel et al. [vBMOW05]. We will, however, analyse the modal logic appropriate to labelled Markov processes.

7.3 Two Examples of Continuous-State Processes

We begin with a simple example, more for introducing terminology than for any intrinsic interest. Imagine a system with two labels $\{a, b\}$. The state space is the real plane, $\mathbf{R}^2$. When the system makes an a-move from state (x_0, y_0) it jumps to (x, y_0) where the probability distribution for x is given by the density $K_\alpha \exp(-\alpha(x - x_0)^2)$, where $K_\alpha = \sqrt{\alpha/\pi}$ is the normalising factor. When it makes a b-move it jumps from state (x_0, y_0) to (x_0, y) where the distribution of y is given by the density function $K_\beta \exp(-\beta(y - y_0)^2)$. The meaning of these densities is as follows. The probability of jumping from (x_0, y_0) to a state with an x-coordinate in the interval $[s, t]$ under an a-move is $\int_s^t K_\alpha \exp(-\alpha(x - x_0)^2)dx$. Note that the probability of jumping to any given point is, of course, 0. In this system the interaction with the environment controls whether the jump is along the x-axis or along the y-axis but the actual extent of the jump is governed by a probability distribution. Interestingly, this system is bisimilar to a one-state system which can always make a or b moves. Thus, from the point of view of an external observer, this system has an extremely simple behaviour. The more complex internal behaviour is not visible to an external observer. The point of a theory of bisimulation that encompasses such systems is to say when systems are equivalent. Of course this example is already familiar from the nonprobabilistic setting; if there is a system in which all transitions are always enabled it will be bisimilar (in the traditional sense) to a system with one state. Bisimulation is a very strong notion and there are much weaker notions that are appropriate in various situations.

Now we consider a system which cannot be reduced to a discrete system. There are three labels $\{a, b, c\}$. Suppose that the state space is $\mathbf{R}$. The state gives the pressure of a gaseous mixture in a tank at a chemical plant. The environment can interact by (a) simply measuring the pressure, or (b) inject some gas into the tank, or (c) pump some gas from the tank. The pressure fluctuates according to some thermodynamic laws depending on the reactions taking place in the tank. With each interaction, the pressure changes according to three different probability density functions, say $f(p_0, p), g(p_0, p)$ and $h(p_0, p)$ respectively, with nontrivial dependence on p_0. In addition, there are two threshold values p_h and p_l. When the pressure rises above p_h the interaction labelled b is disabled, and when the pressure drops below p_l the interaction labelled c is disabled. It is tempting to model this as a three-state system, with the continuous state space partitioned by the threshold values. Unfortunately one cannot assign unique transition probabilities to these sets of states for any choices of f, g and h; only if very special uniformity conditions are obeyed can one do this.

7.4 The Definition of Labelled Markov Processes

In brief, a labelled Markov process can be described as follows. There is a set of states and a set of actions. The system is in a state at a point in time and moves between states. The state to which it moves is governed by the interaction with the environment, and this is indicated by the actions. The system evolves according to a probabilistic law. If the system interacts with the environment through an action it makes a transition to a new state governed by a transition probability distribution. So far, this is essentially the model developed by Larsen and Skou [LS91] in their very important and influential work on probabilistic bisimulation. They specify the transitions by giving, for each label, a probability for going from one state to another. Bisimulation then amounts to matching the moves; this means that both the labels *and the probabilities* must be the same.

In the case of a continuous state space, however, one cannot simply specify transition probabilities from one state to another. In most interesting systems all such transition probabilities would be zero! Instead one must work with probability densities. The precise gadget needed to work with these systems is the stochastic kernel which we have studied in the previous chapters. These will play the role of (probabilistic) transition relations.

A key ingredient in the theory is the transition probability function.

Definition 7.3 A *transition probability function* on a measurable space

(X, Σ) is a function $T : X \times \Sigma \longrightarrow [0,1]$ such that for each fixed $x \in X$, the set function $T(x, \cdot)$ is a (sub)probability measure, and for each fixed $A \in \Sigma$ the function $T(\cdot, A)$ is a measurable function.

One interprets $T(x, A)$ as the probability of the system starting in state x making a transition into one of the states in A. The transition probability is really a *conditional probability*; it gives the probability of the system being in one of the states of the set A after the transition, *given* that it was in the state x before the transition.

It will be convenient to work with *sub-probability* functions, i.e. with functions where $T(x, X) \leq 1$ rather than $T(x, X) = 1$. The mathematical results go through in this extended case, but the stochastic systems studied in the literature are usually only the special cases where $T(x, X)$ is either 1 or 0. In fact what is often done is that a state x with no possibility of making a transition is modelled by having a transition back to itself. For questions concerning which states will eventually be reached (the bulk of the analysis in the traditional literature) this is convenient. If, however, we wish to make contact with the probabilistic verification and process algebra community, we need to make a distinction between a state which can make a transition and one which cannot. In the AI community one often works with observations associated with a state or transition and every state can make a transition with probability 1 with every action. These viewpoints are not really different in any significant way and one can translate back and forth. When we look at MDPs we will adopt the AI conventions.

The mathematical theory of the present chapter requires an analytic space, a concept explained in the next section in order not to disrupt the flow of ideas. Do not worry if you have no idea what they are for the moment. Thus instead of imposing an arbitrary σ-algebra structure on the set of states, we will require that the set of states be an analytic space.

Definition 7.4 A **partial labelled Markov process** with actions $\mathcal{A}$ is a structure $(S, \Sigma, \{\tau_a \mid a \in \mathcal{A}\})$, where S is the set of states, which is assumed to be an analytic space, and Σ is the σ-field on S, and

$$\forall a \in \mathcal{A}, \tau_a : S \times \Sigma \longrightarrow [0,1]$$

is a transition sub-probability function. We are usually interested in the following special case called a **labelled Markov process**. We have a partial labelled Markov process as above and a predicate **Can** on $S \times \mathcal{A}$ such that for every $(x, a) \in$ **Can** we have $\tau_a(x, X) = 1$ and for every $(x, a) \notin$ **Can** we have $\tau_a(x, X) = 0$.

We will fix the actions to be some $\mathcal{A}$ once and for all. The resulting theory is not seriously restricted by this. We will write (S, Σ, τ_a) for partial labelled Markov processes, instead of the more precise $(S, \Sigma, \forall a \in \mathcal{A}.\tau_a)$. In case we are talking about discrete systems, we will use the phrase "labelled Markov chain" rather than "discrete, labelled, Markov process".

7.5 Basic Facts About Analytic Spaces

The basic requirement for the theory to work smoothly is that one should be able to define regular conditional probability densities. As we have seen, regular conditional probability densities are known to exist if we work with Polish spaces. In terms of defining bisimulation, Polish spaces work well, but in characterising bisimulation in terms of a logic we need analytic spaces. The point is that we have to construct quotients of spaces by equivalence relations, but quotients of Polish spaces by equivalence relations are not necessarily Polish. However, analytic spaces have much better stability properties and turn out to be ideally suited for our purposes. The source for the material in this section is Dudley [Dud89] Chapter 13 and Arveson [Arv76] Chapter 3.

Recall that a Polish space is the topological space underlying a *complete, separable* metric space. Any discrete space is Polish. Now a countable product of Polish spaces with the product topology is also Polish. Thus the space $\mathbf{N}^\infty$ is Polish. Any Borel set in a Polish space is the continuous image of $\mathbf{N}^\infty$ but the converse is not true.

Definition 7.5 Let Y be a Polish space. A subset of Y is said to be **analytic** if it is the continuous image of some Polish space.

The following theorem shows that the space $\mathbf{N}^\infty$ plays a special role in the theory.

Theorem 7.6 *Let Y be Polish spaces and let A be a subset of Y. The following six conditions are equivalent:*

(1) A is analytic,
(2) A is the image of some Polish space under a measurable mapping,
(3) A is the continuous image of a Borel subset of a Polish space,
(4) A is the measurable image of a Borel subset of a Polish space,
(5) A is the continuous image of $\mathbf{N}^\infty$,
(6) A is the measurable image of $\mathbf{N}^\infty$.

It is worth noting that the space $\mathbf{N}^\infty$ captures both continuous aspects – for example, every open set in $\mathbf{R}^n$ is the continuous image of it – and discrete aspects, since it is, for example, totally disconnected. The surprising fact is that there are non-Borel analytic sets. The construction is rather complicated but it does not require the axiom of choice. We say that a measurable space is *an analytic space* if it is measurably isomorphic to an analytic set in a Polish space.

The next theorem states that analytic sets, though not always Borel, are always measurable.

Theorem 7.7 *In a Polish space any analytic set is measurable in the σ-field constructed by completing any probability measure.*

Such sets are called *universally measurable.* It is very hard to find nonanalytic universally measurable sets.

The next lemmas are Theorem 3.3.5 of [Arv76] and one of its corollaries. We say that two sets in a Polish space are *separated* if there are contained in disjoint Borel sets. The following result is needed later.

Proposition 7.8 *Two disjoint analytic sets in a Polish space are separated.*

We say that a σ-field *separates points* if, whenever $x \neq y$, there is a measurable set, E, for which $\chi_E(x) \neq \chi_E(y)$. The following powerful theorem and its consequences play a key role in our treatment of the logic.

Theorem 7.9 *Let (X, Σ) be an analytic space and suppose that Σ_0 is a countably generated sub-σ-field of Σ that separates points in X. Then $\Sigma_0 = \Sigma$.*

For our purposes the following corollary is crucial.

Corollary 7.10 *Let X be an analytic space and let $\sim$ be an equivalence relation on X. Assume that there is a sequence $f_1, f_2, \ldots$ of real-valued measurable functions on X such that for all x, y in X we have $x \sim y$ iff for all f_i we have $f_i(x) = f_i(y)$. Then $X/\sim$ is an analytic space.*

This result is the reason why we were forced to work with analytic spaces. All this digression into analytic spaces would be pointless if we did not have the following result.

Theorem 7.11 *Regular conditional probability densities exist on analytic spaces.*

This result can be found in the comprehensive textbook of J. Hoffman-Jørgensen [HJ94b]. In that book a rather more general result is proved and one has to work a little to show that the above theorem is a consequence. The original result is due to Jan Pachl [Pac78] who also proved that no extensions of his result were possible.

7.6 Bisimulation for Labelled Markov Processes

The definition of bisimulation given in our original paper [DEP02] is expressed in categorical terms and is quite subtle and technically complicated. The proof that bisimulation is an equivalence relation depends on a theorem of Edalat [Eda99] which has an intricate proof. Edalat's result was later extended and simplified by Doberkat [Dob03]. In this book, however, we use a definition much closer in spirit to the original Larsen and Skou definition. With this approach it is immediate that bisimulation is an equivalence relation. This approach was discovered later and appeared in our paper on approximation [DGJP03].

The fundamental process equivalence that we consider is *strong probabilistic bisimulation* or just "bisimulation" henceforth. The definition that we use is a slight adaptation of the definition due to Larsen and Skou [LS91], with extra conditions to deal with measure-theoretic issues.

Probabilistic bisimulation means matching the moves and probabilities *exactly* – thus each system must be able to make the same transitions with the same probabilities as the other. Larsen and Skou define a bisimulation relation R as an equivalence relation on the states satisfying the condition that, for each label a, equivalent states have equal probability of making an a-transition to any R-equivalence class of states. In the continuous case, we demand that equivalent states have equal probability of making an a-transition to any union of equivalence classes of states *provided that the union is measurable.*

Instead of talking about sets of equivalence classes we will instead use the notion of *R-closed* sets. Let R be a binary relation on a set S. We say a set $X \subseteq S$ is *R-closed* if $R(X) := \{t | \exists s \in X, sRt\}$ is a subset of X. If R is reflexive, this becomes $R(X) = X$. If R is an equivalence relation, a set is R-closed if and only if it is a union of equivalence classes.

If we want to talk about bisimulation between two different LMPs rather than between the states of a single LMP it will be convenient to explicitly define the notion of *direct sum* of two labelled Markov processes. Then we can define bisimulation on the combined state space. For this it is also convenient to introduce an initial state i for an LMP.

Definition 7.1 Let $\mathcal{S} = (S, i, \Sigma, \tau)$ and $\mathcal{S}' = (S', i', \Sigma', \tau')$ be two labelled Markov processes. The **direct sum** $\mathcal{S} + \mathcal{S}'$ of these processes is a process $\mathcal{U} = (U, u_0, \Omega, \rho)$ where $U = S \uplus S' \cup \{u_0\}$, Ω is the σ-field generated by $\Sigma \cup \Sigma'$, and the transitions are as follows: for all $s \in S$, $s' \in S'$, $\rho_a(s, X \uplus X') = \tau_a(s, X)$ and $\rho_a(s', X \uplus X') = \tau'_a(s', X')$. The initial distribution is given by $\rho_a(u_0, i) = \rho_a(u_0, i') = 1/2$.

This construction is purely formal and is only used in order to define a relation on the common state space.

Definition 7.2 Let $\mathcal{S} = (S, i, \Sigma, \tau)$ be a labelled Markov process. An equivalence relation R on S is a **bisimulation** if whenever sRs', with $s, s' \in S$, we have that for all $a \in \mathcal{A}$ and every R-closed measurable set $X \in \Sigma$, $\tau_a(s, X) = \tau_a(s', X)$. Two states are bisimilar if they are related by a bisimulation relation.

Let $\mathcal{S} = (S, i, \Sigma, \tau)$ and $\mathcal{S}' = (S', i', \Sigma', \tau')$ be a pair of labelled Markov process. $\mathcal{S}$ and $\mathcal{S}'$ are bisimilar if there is a bisimulation relation on some process $\mathcal{U}$, of which $\mathcal{S}$ and $\mathcal{S}'$ are direct summands, relating i and i' in $\mathcal{U}$.

Alternately, bisimulation on the states of a labelled Markov process can be viewed as the maximum fixed point of the following (monotone) functional F on the lattice of equivalence relations on $(S \times S, \subseteq)$:

$$s\,F(R)\,t \text{ if for all } a \in \mathcal{A}, \text{and all } R\text{-closed } C \in \Sigma, \tau_a(s, C) \leq \tau_a(t, C).$$

The intuition of this definition is that the relation R relates those states that can be "lumped" together. Bisimulation is the largest such relation. In fact the notion of bisimulation (for systems with a single action and no nondeterminism) was known in the queuing theory community [KS60] under the term "lumpability". With regard to bisimulation on processes, note that we do not require $\mathcal{U}$ to be exactly $\mathcal{S} + \mathcal{S}'$ but rather a sum of a number of processes, of which are $\mathcal{S}$ and $\mathcal{S}'$. The reason for this is that transitivity of bisimulation would not follow in any obvious way with $\mathcal{U}$ being exactly the direct sum. However, the logical characterisation of bisimulation below will allow us to restrict ourselves to only considering $\mathcal{U} = \mathcal{S} + \mathcal{S}'$ in the above definition.

With our definition of bisimulation the following proposition is easy.

Proposition 7.12 *Bisimulation is an equivalence relation.*

Proof. Bisimulation is obviously reflexive and symmetric. For transitivity, consider two bisimulations R_1 and R_2 on a single process $\mathcal{S} = (S, i, \Sigma, \tau)$. Let R be the transitive closure of $R_1 \cup R_2$. Then every measurable R-closed set is also R_i-closed, $i = 1, 2$, and it follows easily that R is a bisimulation on $\mathcal{S}$.

In the case of bisimulation between processes, let R_1 be a bisimulation between $\mathcal{S}$ and $\mathcal{S}'$ through process $\mathcal{U}_1 = \mathcal{S} + \mathcal{S}' + \mathcal{T}$ (for some $\mathcal{T}$) and R_2 a bisimulation between $\mathcal{S}'$ and $\mathcal{S}''$ through process $\mathcal{U}_2 = \mathcal{S}' + \mathcal{S}'' + \mathcal{T}'$(for some $\mathcal{T}'$). Then construct the direct sum $\mathcal{U} = \mathcal{U}_1 + \mathcal{U}_2$. Consider the following relations in $\mathcal{U} \times \mathcal{U}$:

- Let $\mathtt{I}$ be the symmetric and reflexive closure of the relation in $\mathcal{U} \times \mathcal{U}$ that relates copies of states from $\mathcal{S}'$ in $\mathcal{U}_1$ to copies of the same states in $\mathcal{U}_2$.
- Let $E_1 = R_1 + \mathtt{id}_{\mathcal{U}_2}$, $E_2 = \mathtt{id}_{\mathcal{U}_1} + R_2$

It is easy to show that I, E_1, E_2 are bisimulations on $\mathcal{U}$. Considering the transitive closure of $I \cup E_1 \cup E_2$ as in the above proof yields the required result. □

A functional analogue of the concept of bisimulation is a function called a zigzag.

Definition 7.13 A function f from (S, Σ, τ_a) to (S', Σ', τ'_a) is *a zigzag* if it satisfies the properties:

(1) f is surjective;
(2) f is measurable;
(3) $\forall a \in \mathcal{A}, s \in S, \sigma' \in \Sigma', \;\; \tau_a(s, f^{-1}(\sigma')) = \tau'_a(f(s), \sigma')$.

Note that the states are mapped forward by f but the measurable sets are mapped backwards by f^{-1}; this is why the name "zigzag" is used. The name "zigzag" comes from modal logic [Pop94]. Requiring f to be surjective allows us to avoid introducing initial states and worrying about reachable states. One can immediately check that the identity function is a zigzag. We had originally defined bisimulation as a span of zigzags [DEP02]. In the coalgebraic approach to probabilistic systems [RdV97] zigzags appear naturally as the coalgebra homomorphisms.

7.7 A Logical Characterisation of Bisimulation

One can define a simple modal logic and prove that two states are bisimilar if and only if they satisfy exactly the same formulas. Indeed for finite-state processes one can decide whether two states are bisimilar and effectively construct a distinguishing formula in case they are not [DEP02]. What was a surprise is that one does not need infinite branching – even though we have infinite branching – or negation or even any kind of limited negative construct. This is in striking contrast with what was known for the discrete case [LS91] where a strong finite branching assumption was made and negative constructs were needed in the logic.

As before we assume that there is a fixed set of "actions" $\mathcal{A}$. The logic is called $\mathcal{L}$ and has the following syntax:

$$\mathsf{T} \mid \phi_1 \wedge \phi_2 \mid \langle a \rangle_q \phi$$

where a is an action and q is a rational number. This is the basic logic with which we establish the logical characterisation. In later sections we will work with this logic augmented with disjunction, $\mathcal{L}_\vee$:

$$\mathcal{L} \mid \phi_1 \vee \phi_2.$$

Definition 7.3 We define the **depth** of formulas inductively as follows:

$$\begin{aligned} depth(\mathsf{T}) &= 0 \\ depth(\phi \wedge \psi) &= max(depth(\phi), depth(\psi)) \\ depth(\phi \vee \psi) &= max(depth(\phi), depth(\psi)) \\ depth(\langle a \rangle_r \phi) &= depth(\phi) + 1 \end{aligned}$$

Given a labelled Markov process $\mathcal{S} = (S, i, \Sigma, \tau)$ we write $s \models \phi$ to mean that the state s satisfies the formula ϕ. The definition of the relation $\models$ is given by induction on formulas. The definition is obvious for the propositional constant T, conjunction and, when we introduce $\mathcal{L}_\vee$, disjunction as well. We say $s \models \langle a \rangle_q \phi$ if and only if $\exists X \in \Sigma.(\forall s' \in X.s' \models \phi) \wedge (\tau_a(s, X) > q)$. In other words, the process in state s can make an a-move to a state, that satisfies ϕ, with probability strictly greater than q[2]. We write $[\![\phi]\!]_{\mathcal{S}}$ for the set $\{s \in S \mid s \models \phi\}$, and $\mathcal{S} \models \phi$ if $i \models \phi$. We often omit the subscript when no confusion can arise.

The main theorem relating $\mathcal{L}$ and bisimulation is the logical characterisation of bisimulation. This was proved in [DEP98; DEP02]. The present

[2] In our earlier work we had used $\geq$ instead of $>$.

proof is adapted from that to our present relational presentation of bisimulation.

First we need a straightforward but important lemma.

Lemma 7.1 *For any formula ϕ in $\mathcal{L}_\vee$ the set $[\![\phi]\!]$ is measurable.*

Proof. We proceed by structural induction on ϕ. The base case corresponding to T is trivial since $S \in \Sigma$. Conjunction and disjunction are trivial because, by definition, a σ-field is closed under intersection and union. Finally, we have $[\![\langle a\rangle_q\phi]\!] = \tau_a(\cdot, [\![\phi]\!])^{-1}([q,1]) \in \Sigma$. To justify this first note that, by hypothesis, $[\![\phi]\!] \in \Sigma$ by the inductive hypothesis so $\tau_a(s, [\![\phi]\!])$ is meaningful. Secondly, τ_a is a measurable function in its first argument and intervals are Borel-measurable so the set $[\![\langle a\rangle_q\phi]\!]$ is the inverse image of a measurable set by a measurable function. □

The fact that bisimilar states satisfy the same formulas is an easy proposition.

Proposition 7.14 *Let R be a bisimulation on $\mathcal{S}$, if sRs' then s and s' satisfy the same formulas.*

Proof. We prove the proposition by induction on the structure of formulas. The cases of T and conjunction are trivial. Now assume the implication is true for ϕ, i.e., for every pair of R-related states either both satisfy ϕ or neither of them does. This means that the set $[\![\phi]\!]$ is R-closed, and by Lemma 7.1 is measurable.

Since R is a bisimulation, $\tau_a(s, [\![\phi]\!]) = \tau_a(s', [\![\phi]\!])$ for all $a \in \mathcal{A}$. So s and s' satisfy the same modal formulas of the form $\langle a\rangle_q\phi$. □

The converse depends on some remarkable special properties of analytic spaces. The first is a very strong "rigidity" property: it says that if one has a sub-σ-algebra, say Λ of an analytic space Σ and Λ separates points[3] (so it is not too small) and countably generated (so it is not too large) then Λ is all of Σ. The second says that if one takes the quotient of an analytic space in a "reasonable" way then the result is analytic.

Lemma 7.2 *Let $(X, \mathcal{B})$ be an analytic space and let $\mathcal{B}_0$ be a countably generated sub-σ-field of $\mathcal{B}$ which separates points in X. Then $\mathcal{B}_0 = \mathcal{B}$.*

Lemma 7.3 *Let X be an analytic space and let $\sim$ be an equivalence relation in X. Assume there is a sequence $f_1, f_2, \ldots$ of real-valued Borel*

[3] This means that for every pair of distinct points there is a measurable set that contains one but not the other.

functions on X such that for any pair of points x, y in X one has $x \sim y$ iff $f_n(x) = f_n(y)$ for all n. Then $X/_\sim$ is an analytic space.

These lemmas are from Chapter 3 of *Invitation to C^* Algebras* by Arveson [Arv76].

One can see that Lemma 7.3 is ideally suited to our purposes. We have a natural collection of such functions that we want to use: the formulas of $\mathcal{L}$, or more precisely the characteristic functions of the sets $[\![\phi]\!]$ for all the formulas in $\mathcal{L}$. Recall that there are only countably many formulas in $\mathcal{L}$.

We introduce a *logical* equivalence relation between states.

Definition 7.15 Let $\mathcal{S} = (S, i, \Sigma, \tau)$ be an LMP and s and t be states of $\mathcal{S}$. We say that $s \approx t$ if

$$\forall \phi \in \mathcal{L}\; s \models \phi \iff t \models \phi.$$

Our main goal is to show that $\approx$ is a bisimulation relation.

We first show that there is a zigzag from any system $\mathcal{S}$ to its quotient under $\approx$. If (S, Σ) is a measurable space, the quotient $(S/_\approx, \Sigma_\approx)$ is defined as follows. $S/_\approx$ is the set of all equivalence classes. Then the function $q : S \to S/_\approx$ which assigns to each point of S the equivalence class containing it maps onto $S/_\approx$, and thus determines a σ-algebra structure on $S/_\approx$: by definition a subset E of $S/_\approx$ is a measurable set if $q^{-1}(E)$ is a measurable set in (S, Σ).

Proposition 7.16 *Let (S, Σ, τ_a) be an LMP. We can define ρ_a so that the canonical projection q from (S, Σ, τ_a) to $(S/_\approx, \Sigma_\approx, \rho_a)$ is a zigzag morphism.*

In order to prove this proposition we need a couple of lemmas in addition to Lemmas 7.2 and 7.3.

The first lemma just says that the transition probabilities to sets of the form $[\![\phi]\!]$ are completely determined by the formulas.

Lemma 7.4 *Let (S, Σ, τ_a) and (S', Σ', τ'_a) be two LMPs. Then for all formulas ϕ and all pairs (s, s') such that $s \approx s'$, we have $\tau_a(s, [\![\phi]\!]_S) = \tau'_a(s', [\![\phi]\!]_{S'})$.*

Proof. Suppose that the equation does not hold. Then, say, for some ϕ, $\tau_a(s, [\![\phi]\!]_S) < \tau'_a(s', [\![\phi]\!]_{S'})$. We choose a rational number q between these values. Now it follows that $s' \models \langle a \rangle_q \phi$ but $s \not\models \langle a \rangle_q \phi$, which contradicts the assumption that s and s' satisfy all the same formulas. □

The final result that we need is Prop. 2.10 which tells us that when two measures agree on a π-system they will agree on the generated σ-algebra.

This is again just what we need; the formulas are closed under conjunction so the collection of sets $[\![\phi]\!]$ is a π-system.

Proof of Proposition 7.16: We first show that $S/_{\approx}$ is an analytic space. Let $\{\phi_i | i \in \mathbf{N}\}$ be the set of all formulas. We know that $[\![\phi_i]\!]_S$ is a measurable set for each i. Therefore the characteristic functions $\chi_{\phi_i} : S \to \{0,1\}$ are measurable functions. Moreover we have

$$x \approx y \text{ iff } (\forall i \in \mathbf{N}.\ x \in [\![\phi_i]\!]_S \iff y \in [\![\phi_i]\!]_S) \text{ iff } (\forall i \in \mathbf{N}.\ \chi_{\phi_i}(x) = \chi_{\phi_i}(y)).$$

It now follows by Lemma 7.3 that $S/_{\approx}$ is an analytic space.

Let $\mathcal{B} = \{q([\![\phi_i]\!]_S) : i \in \mathbf{N}\}$. We show that $\sigma(\mathcal{B}) = \Sigma_{\approx}$. We have inclusion since $\mathcal{B} \subseteq \Sigma_{\approx}$; indeed, for any $q([\![\phi_i]\!]_S) \in \mathcal{B}$, $q^{-1}q([\![\phi_i]\!]_S) = [\![\phi_i]\!]_S$ which is in Σ by Lemma 7.1. Now $\sigma(\mathcal{B})$ separates points in $S/_{\approx}$, for if x and y are different states of $S/_{\approx}$, take states $x_0 \in q^{-1}(x)$ and $y_0 \in q^{-1}(y)$. Then since $x_0 \not\approx y_0$, there is a formula ϕ such that x_0 is in $[\![\phi]\!]_S$ and y_0 is not. It means that

$$\forall s \in q^{-1}(x), s \in [\![\phi]\!]_S \text{ and } \forall t \in q^{-1}(y), t \notin [\![\phi]\!]_S$$

so that x is in $q[\![\phi]\!]_S$, whereas y is not. Since $\sigma(\mathcal{B})$ is countably generated, it follows by Lemma 7.2, that $\sigma(\mathcal{B}) = \Sigma_{\approx}$.

We are now ready to define $\rho_a(t,\cdot)$ over $\Sigma_{\approx}$ for $t \in S/_{\approx}$. We define it so that $q : S \to S/_{\approx}$ is a zigzag (recall that q is measurable and surjective by definition), i.e., for any $B \in \Sigma_{\approx}$ we put

$$h_a(t,B) = \tau_a(s, q^{-1}(B)),$$

where $s \in q^{-1}(t)$. Clearly, for a fixed state s, $\tau_a(s, q^{-1}(\cdot))$ is a subprobability measure on $\Sigma_{\approx}$. We now show that the definition does not depend on the choice of s in $q^{-1}(t)$ for if $s, s' \in q^{-1}(t)$, we know that $\tau_a(s, q^{-1}(\cdot))$ and $\tau_a(s', q^{-1}(\cdot))$ agree over $\mathcal{B}$ again by the fact that $q^{-1}q([\![\phi_i]\!]_S) = [\![\phi_i]\!]_S$ and by Lemma 7.4. So, since $\mathcal{B}$ is closed under the formation of finite intersections we have, from Prop. 2.10, that $\tau_a(s, q^{-1}(\cdot))$ and $\tau_a(s', q^{-1}(\cdot))$ agree on $\sigma(\mathcal{B}) = \Sigma_{\approx}$.

It remains to prove that for a fixed Borel set B of $\Sigma_{\approx}$, $\rho_a(\cdot, B) : S/_{\approx} \to [0,1]$ is a Borel measurable function. Let A be a Borel set of $[0,1]$. Then $\rho_a(\cdot,B)^{-1}(A) = q[\tau_a(\cdot, q^{-1}(B))^{-1}(A)]$; we know that $\sigma = \tau_a(\cdot, q^{-1}(B))^{-1}(A)$ is Borel since it is the inverse image of A under a Borel measurable function. Now we have that $q(\sigma) \in \Sigma_{\approx}$, since $q^{-1}q(\sigma) = \sigma$; indeed, if $s_1 \in q^{-1}q(\sigma)$, there exists $s_2 \in \sigma$ such that $q(s_1) = q(s_2)$, and we have just proved above that then the $\tau_a(s_i, q^{-1}(\cdot))$'s must agree. So if

$\tau_a(s_i, q^{-1}(B)) \in A$ for $i = 2$, then it is also true for $i = 1$, so $s_1 \in \sigma$ as wanted. So $\rho_a(\cdot, B)$ is measurable. This concludes the proof that $S/_{\approx}$ is an LMP and q a zigzag. □

Theorem 7.17 *Let (S, i, Σ, τ) be a labelled Markov process. Two states $s, s' \in S$ are bisimilar if and only if they satisfy the same formulas of $\mathcal{L}$.*

Proof. ⇒: This is just Prop. 7.14.

⇐: We will use Prop. 7.16 to show that the relation $\approx$ defined on the states of $\mathcal{S}$ is in fact a bisimulation relation. The key facts are

(1) $B \in \Sigma_{\approx}$ if and only if $q^{-1}(B) \in \Sigma$,
(2) $\forall s \in S, B \in \Sigma_{\approx}.\rho(q(s), B) = \tau(s, q^{-1}(B))$.

Let $X \in \Sigma$ be $\approx$-closed. Then we have $X = q^{-1}(q(X))$ and hence $q(X) \in \Sigma_{\approx}$. Now if $s \approx s'$, then $q(s) = q(s')$, and $\tau_a(s, X) = \rho_a(f(s), f(X)) = \tau_a(s', X)$, and hence $\approx$ is a bisimulation. □

The strategy of the proof is worth recalling since it can be used in several other situations. We use the logic to deduce that the transition probabilities coincide on sets definable in the logic; the fact that these sets form a π-system is important. At first sight, this seems far from being all measurable sets. However, the unique structure theorem forces this to extend to all sets.

As a corollary, we deduce that the closure ordinal of the functional F defining bisimulation is ω. This can be skipped for now but it is interesting in connection with later results on approximation and the metric.

Corollary 7.18 *Define a family of relations $R_i \subseteq LMP \times LMP$ as follows:*

$$\begin{aligned} R_0 &= \mathbf{LMP} \times \mathbf{LMP} \\ R_{i+1} &= F(R_i) \textit{ for } F \textit{ as defined in discussion of Definition 7.2} \\ R &= \cap_i R_i \end{aligned}$$

pRq if and only if p is bisimilar to q.

Proof. The bisimulation relation is contained in R_0; hence by induction on i and by using $R_{i+1} = F(R_i)$, and the fact that bisimulation is a fixed point of the monotone functional F, we get that the bisimulation relation is contained in R_i for all i, and hence is contained in $R = \cap_i R_i$.

Now we need to prove that pRq implies p is bisimilar to q. We prove by induction on i that $[\![\phi]\!]$ is the union of R_i equivalence classes for $i \geq d(\phi)$.

Formally, sR_it implies for all formulas ϕ of depth i or less, $s \models \phi$ if and only if $t \models \phi$. The base case is trivial.

Inductive step: Let $sR_{i+1}t$. Consider a formula $\langle a \rangle_r \phi$ such that $s \models \langle a \rangle_r \phi$. Thus $\tau_a(s, [\![\phi]\!]) > r$. But since $[\![\phi]\!]_{\mathcal{S}}$ is an R_i equivalence class and is measurable, $\tau_a(t, [\![\phi]\!]_{\mathcal{S}}) = \tau_a(s, [\![\phi]\!]_{\mathcal{S}}) > r$. □

Chapter 8

Metrics for Labelled Markov Processes

The main concept that we have used in the study of labelled Markov processes is bisimulation, which is a behavioural *equivalence* between processes. Though bisimulation has proven to be a central idea in concurrency theory, there is a serious objection to using it for probabilistic transition systems or any other type of transition system – such as real-time systems – where real numbers play a role. The point is that bisimulation cannot distinguish between two systems that differ by a *small* amount in the real-valued parameters two systems that are completely different. What is needed is a *quantitative* measure of *how* different systems are. This was first emphasised by Jou and Smolka [JS90].

A natural idea is to define a notion of approximate bisimulation by saying that the relevant transition probabilities should be close: for example, they have to be within some given small real number ϵ. Unfortunately this does not even yield an equivalence relation. This is, however, not a silly idea and can be made to work through the notion of a *uniformity*. We will not pursue this idea here, instead we will seek a direct quantitative analogue of the idea of an equivalence relation. What is needed is a *metric* or, more precisely, a pseudometric.

Definition 8.1 A **pseudometric** on a set X is a nonnegative real-valued function $d : X \times X \longrightarrow \mathbf{R}$ satisfying:

(1) $\forall x \in X \; d(x, x) = 0$
(2) $\forall x, y \in X \; d(x, y) = d(y, x)$
(3) $\forall x, y, z \in X \; d(x, y) \leq d(x, z) + d(y, z)$.

If, in addition, we require that $\forall x, y \in X \; d(x, y) = 0 \Rightarrow x = y$, we get a **metric**.

Most of the time we will deal with pseudometrics, i.e. we will allow two distinct points to be at zero distance. Given a pseudometric one can define an associated equivalence relation by deeming two points to be equivalent if they are at zero distance; it is an easy exercise to see that this is indeed an equivalence relation. Our goal will be to develop a notion of pseudometric between probabilistic transition systems (or states of a probabilistic transition system) such that the induced equivalence is bisimulation. Thus the pseudometric will be a natural quantitative extension of the notion of bisimulation.

Before we continue we should understand why the above axioms for a pseudometric are chosen. Suppose we informally use the phrasing "x and y are close" to mean that $d(x, y)$ is "small", then the first axiom says that every point is close to itself, the second says that if x is close to y then y is close to x and the third says that if x is close to z and z is close to y then x is close to y. These exactly correspond to the axioms for an equivalence relation. The requirement for a metric says, very roughly speaking, that the only point close to a given point is the point itself. Henceforth we will just say "metric" rather than the more correct "pseudometric", unless there is a possibility of genuine confusion.

8.1 From Bisimulation to a Metric

The basic intuition behind our metrics is as follows. In view of our earlier results on the logical characterisation of bisimulation, we know that if two processes are not bisimilar there will be a formula that distinguishes them. We measure the distance between processes in terms of the smallest formula required to distinguish them. If the formula is very large then only a long sequence of observations will distinguish the processes. This view, as stated, does not take into account the fact that the processes might differ immediately but do so with probabilities that are very close. There are thus two dimensions along which processes could differ: their future behaviour and the immediate transition probabilities. In order to handle the latter possibility we need some notion of formulas that are "close". We handle this semantically by introducing a quantitative analogue of logical formula. The technical development of these intuitions is based on an idea expounded by Kozen [Koz85] to generalise logic to handle probabilistic phenomena. Later we will show how the metric can be viewed as a fixed point in a manner very similar to bisimulation.

We recapitulate Kozen's ideas using the following table.

Classical logic	Generalisation
Truth values $\{0, 1\}$	Interval $[0, 1]$
Propositional function	Measurable function
State	Measure
The satisfaction relation $\models$	Integration $\int$

Just as the satisfaction relation, $\models$, links states and formulas to give truth values, so the integral links measures (generalised states) with measurable functions (generalised formulas) to give real numbers (generalised truth values).

Following these intuitions, we consider a class $\mathcal{F}$ of functions that assign a value in the interval $[0, 1]$ to states of a process. These functions are inspired by the formulas of $\mathcal{L}$ – the result of evaluating these functions at a state corresponds to a quantitative measure of the extent to which the state satisfies a formula of $\mathcal{L}$. The identification of this class of functions is a key aspect of the present development and turns out to be closely related to value functions for MDPs. The definition of these functions leads to a metric d:

$$d(\mathcal{P}, \mathcal{Q}) = \sup\{|f(p_0) - f(q_0)| \mid f \in \mathcal{F}\}.$$

In Section 8.3 we will formalise the above intuitions to define a family of metrics $\{d^c \mid c \in (0, 1]\}$. These metrics support the spectrum of possibilities of relative weighting of the two factors that contribute to the distance between processes: the complexity of the functions distinguishing them versus the amount by which each function distinguishes them. The metric d^1 captures only the differences in the probabilities; probability differences at the first transition are treated on par with probability differences that arise very deep in the evolution of the process. In contrast, the metrics d^c for $c < 1$ give more weight to the probability differences that arise earlier in the evolution of the process, i.e. differences identified by simpler functions. As c approaches 0, the future gets discounted more.

As is usual with metrics, the actual numerical values of the metric are less important than properties like the significance of zero distance, relative distance of processes, contractivity and the notion of convergence (i.e. the *topology*). If the discount factor c is strictly between 0 and 1 none of these properties are affected. If, however the discount factor is 1 – corresponding to no discount – then the notion of convergence is drastically affected.

Example 8.1 Consider the process $\mathcal{P}$ with two states, and a transition going from the start state to the other state with probability p. Let $\mathcal{Q}$ be a similar process, with the probability q. Then in Section 8.3, we show that $d^c(\mathcal{P},\mathcal{Q}) = c|p-q|$. Now if we consider $\mathcal{P}'$ with a new start state, which makes a b transition to $\mathcal{P}$ with probability 1, and similarly $\mathcal{Q}'$ whose start state transitions to $\mathcal{Q}$ on $\mathbf{b}$ with probability 1, then $d^c(\mathcal{P}',\mathcal{Q}') = c^2|p-q|$, showing that the next step is discounted by c.

Each of these metrics agree with bisimulation:

$$d^c(\mathcal{P},\mathcal{Q}) = 0, \text{ iff } \ \mathcal{P} \ \text{ and } \ \mathcal{Q} \ \text{ are bisimilar.}$$

8.2 A Real-Valued Logic on Labelled Markov Processes

In this section we present an alternate characterisation of probabilistic bisimulation using functional expressions – which define functions into the reals – instead of the logic $\mathcal{L}$. We define a set of functions that are sufficient to characterise bisimulation. It is worth clarifying our terminology here. We define a set of *functional expressions* or *real-valued formulas* by giving an explicit syntax. A functional expression becomes a function when we interpret it in a system. Thus we may loosely say "the same function" when we move from one system to another. What we really mean is the "same functional expression"; obviously it cannot be the same function when the domains are different. This is no different from having syntactically defined formulas of some logic which become boolean-valued functions when they are interpreted.

Definition 8.1 For each $c \in (0,1]$, we consider a family $\mathcal{F}^c$ of functional expressions generated by the following grammar.

$$f := \mathbf{1} \mid \mathbf{1} - f \mid \langle a \rangle f \mid \min(f_1, f_2) \mid \sup_{i \in \mathbf{N}} f_i \mid f \ominus q,$$

where q is a rational. $\mathcal{F}^c_+$ is the sub-collection of $\mathcal{F}^c$ that does not use the negation functional $\mathbf{1} - f$ and $\mathcal{F}in^c_+$ is the sub-collection of $\mathcal{F}^c_+$ that uses finite sup only.

The interpretation is as follows. Let $\mathcal{S} = (S, s_0, \Sigma, \tau)$ be a labelled Markov process. We write $f_{\mathcal{S}} : S \longrightarrow [0,1]$ for the interpretation of $f \in \mathcal{F}^c$

on $\mathcal{S}$ and drop the subscript when no confusion can arise. Let $s \in S$. Then

$$\begin{aligned}
&\mathbf{1}(s) = 1,\\
&(\mathbf{1} - f)(s) = 1 - f(s),\\
&\langle a\rangle f(s) = c\int_S f(t)\tau_a(s, dt),\\
&(f \ominus q)(s) = \max(f(s) - q, 0),
\end{aligned}$$

and min and sup are defined in the obvious way.

In the interpretation of $\langle a\rangle f$, c refers to the constant in $\mathcal{F}^c$; this is the only place where an explicit mention of c occurs. We use $\langle a\rangle^n f$ to represent $\langle a\rangle \cdots \langle a\rangle f$ where $\langle a\rangle$ appears n times.

One can think of these functional expressions as being associated with the logical connectives of $\mathcal{L}$ in the following way. T is represented by the functional $\mathbf{1}$ and conjunction by min. The role of the connective $\langle a\rangle_q$ is split up into two expressions: $\langle a\rangle f$, which intuitively corresponds to prefixing, and $f \ominus q$, which captures the "greater than q" idea.

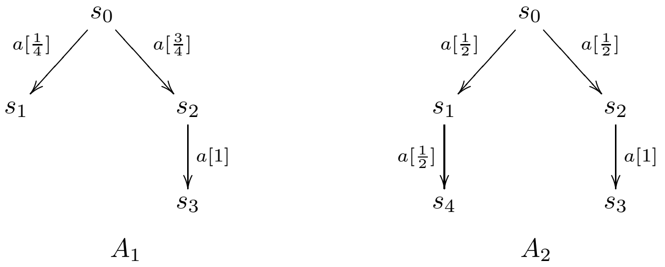


Fig. 8.1 Labelled Markov chains

Example 8.2 Consider the finite processes A_1 and A_2 in Figure 8.1. The functional expression $(\langle a\rangle\mathbf{1})$ of $\mathcal{F}^c$ evaluates to c at states s_0, s_2 of both A_1 and A_2; it evaluates to 0 at states s_1, s_3 of A_1 and s_3, s_4 of A_2, and it evaluates to $c/2$ at state s_1 of A_2. The functional expression $(\langle a\rangle.\langle a\rangle\mathbf{1})$ evaluates to $3c^2/4$ at states s_0 of A_1, A_2 and to 0 elsewhere. The functional expression $(\langle a\rangle(\langle a\rangle\mathbf{1} \ominus \frac{c}{2}))$ evaluates to $3c^2/8$ at state s_0 of A_1 and to $c^2/4$ at state s_0 of A_2. This example shows the need for the connective $\ominus$ in the functional expressions. Without it there would be no way of distinguishing these two. Note, however, that this example relies on the fact that one can have subprobability distributions associated with a labelled transition. If we insisted that all probability distributions had to be normalised then functional expressions without $\ominus$ would suffice [dAHM03].

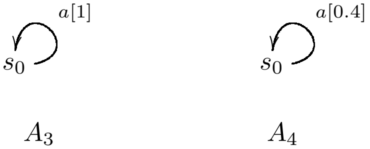


Fig. 8.2 Labelled Markov chains

Example 8.3 Consider the finite process A_3 in Figure 8.2 and functionals of $\mathcal{F}^c$. A functional expression of the form $(\langle a \rangle^n.\mathbf{1})$ evaluates to c^n at state s_0. At state s_0 of process A_4 the same functional expression evaluates to $(c \times 0.4)^n$.

There is a close relation between the values of these functional expressions and simulation (and bisimulation). The results in the rest of this section are from Desharnais's thesis [Des99]; the proofs are provided for completeness and can be safely skipped without loss of continuity.

A routine induction on the structure of the functional expression $f \in \mathcal{F}^c_+$, shows:

Lemma 8.1 *If $\mathcal{S}$ is simulated by $\mathcal{S}'$, then $\forall s, s'$ such that s and s' are related by the simulation relation we have*

$$(\forall f \in \mathcal{F}^c_+)\ [f_{\mathcal{S}}(s) \leq\ f_{\mathcal{S}'}(s')].$$

The next several lemmas and their corollaries – from Lemma 8.2 to Corollary 8.2 – are aimed at proving that the functional expressions characterise bisimulation. The proof below uses our earlier results on the logical characterisation of bisimulation. It is also possible to proceed directly, essentially using the same techniques readapted to functional expressions. However, the proof below also shows how the functional expressions and the logical formulas are related.

For any finite process $\mathcal{P}$ and any formula, there is a functional from $\mathcal{F}^c_+$ which distinguishes between states of $\mathcal{P}$ that do or do not satisfy the formula. This functional furthermore gives a zero value to any state of any process that does not satisfy the formula.

Lemma 8.2 *Given $\phi \in \mathcal{L}_\vee$, a finite process $\mathcal{P}$, and $c \in (0,1]$, there is a functional expression $f \in \mathcal{F}^c_+$ such that*

(1) $\forall p \in P$ we have $f_{\mathcal{P}}(p) > 0$ iff $p \models_{\mathcal{P}} \phi$;
(2) for any state s of any labelled Markov process $\mathcal{S}$, we have $f_{\mathcal{S}}(s) > 0 \Rightarrow s \models_{\mathcal{S}} \phi$.

Proof. The proof is by induction on the structure of ϕ. The key case is $\phi = \langle a \rangle_q \psi$, let g be the functional expression corresponding to ψ yielded by induction. Let $x = \min\{g(t) \mid t \in [\![\psi]\!]_{\mathcal{P}}\}$. By induction hypothesis, $x > 0$. Recall that a constant function $1 - q$ on processes can be obtained with the functional $\mathbf{1} \ominus q$: consequently we can legitimately use the notation $\min(g, x)$ to mean $\min(g, \mathbf{1} \ominus (1 - x))$. Consider the functional expression f given by $(\langle a \rangle \min(g, x)) \ominus cxq$. For all $t \in [\![\psi]\!]_{\mathcal{P}}$, $\min(g, x)(t) = x$. Now for any state $p \in P$,

$$\langle a \rangle \min(g, x)(p) = cx \sum_{t \in [\![\psi]\!]_{\mathcal{P}}} \tau_a(p, t) = cx\tau_a(p, [\![\psi]\!]_{\mathcal{P}}).$$

Now the last expression is $> cxq$ if and only if $p \in [\![\langle a \rangle_q . \psi]\!]_{\mathcal{P}}$. Thus f satisfies the first condition.

The second condition holds because for any $s \in S$, $\langle a \rangle \min(g, x)(s) \leq cx\tau_a(s, [\![\psi]\!]_{\mathcal{S}})$, so if $s \not\models \phi$ then $\tau_a(s, [\![\psi]\!]_{\mathcal{S}}) \leq q$ and hence $f(s) = 0$. □

Note that if the formula is finite, then the corresponding functional lies in $\mathcal{F}in^c_+$. The previous lemma can be partially extended to arbitrary labelled Markov processes. In this case, the functional corresponding to a formula does not work for *every* state of the process. The functional will depend on the states and the formula must be finite. We omit the latter proofs since they require the approximation results; the complete proofs appear in [DGJP04].

In order to proceed we need some finiteness properties. The lemma below, Lemma 8.3 says that the truth of formulas can be witnessed by finite sub-processes.

Definition 8.2 $\mathcal{S} = (S, s_0, \Sigma, \tau)$ is a sub-process of $\mathcal{S}' = (S', s'_0, \Sigma', \tau')$ if $(S, \Sigma) \subseteq (S', \Sigma')$ (this means that the inclusion map $S \subseteq S'$ is measurable), $s_0 = s'_0$ and for every $a \in \mathcal{A}$, $s \in S$, $X \in \Sigma$ we have $\tau_a(s, X) \leq \tau'_a(s, X)$.

Thus, a sub-process has fewer states and lower probabilities than the original process.

Lemma 8.3 *Let $\mathcal{P}$ be a labelled Markov chain, $p \in P$ and $\phi \in \mathcal{L}_\vee$ such that $p \models_{\mathcal{P}} \phi$. Then there exists a finite sub-process of $\mathcal{P}$, $\mathcal{Q}^p_\phi$, such that $p \in Q^p_\phi$, and $p \models_{\mathcal{Q}^p_\phi} \phi$.*

Proof. The proof is by induction on ϕ. For T, the one state process containing p suffices. For $\phi = \phi_1 \wedge \phi_2$, we take the union of the finite processes, $\mathcal{Q}^p_{\phi_1}, \mathcal{Q}^p_{\phi_2}$ given by the induction hypothesis, which ensures that $p \models_{\mathcal{Q}^p_\phi} \phi_1 \wedge \phi_2$. For disjunction, $\bigvee_{i=1}^{\infty} \phi_i$, we take $\mathcal{Q}^p_{\phi_1}$ (or any other $\mathcal{Q}^p_{\phi_i}$).

Let $p \models_{\mathcal{P}} \langle a \rangle_r \psi$. Then, since $\tau_a(p, [\![\psi]\!]_{\mathcal{P}}) > r$, there is a finite subset $U = \{p_1, \ldots, p_n\} \subseteq [\![\psi]\!]_{\mathcal{P}}$, such that $\tau_a(p, U) > r$. The required finite process, $\mathcal{Q}^p_{\langle a \rangle_r . \psi}$, is now constructed by taking the unions of the finite processes, $\mathcal{Q}^{p_1}_\psi, \ldots, \mathcal{Q}^{p_n}_\psi$, adding state p and transitions from p to p_i for $i = 1 \ldots\ n$. □

Proposition 8.1 *Given $\phi \in \mathcal{L}_\vee$, a labelled Markov chain $\mathcal{P}$, $c \in (0,1]$ and a state $p \in P$, if $p \models_{\mathcal{P}} \phi$, then there exists $f \in \mathcal{F}in^c_+$ such that*

(1) $f_{\mathcal{P}}(p) > 0$ and
(2) for any state s of any labelled Markov process $\mathcal{S}$, we have $f_{\mathcal{S}}(s) > 0 \Rightarrow s \models_{\mathcal{S}} \phi$.

Proof. Let p be a state in P such that $p \models_{\mathcal{P}} \phi$. By Lemma 8.3, there is a finite sub-process $\mathcal{Q}^p_\phi$ of $\mathcal{P}$ such that $p \models_{\mathcal{Q}^p_\phi} \phi$. By Lemma 8.2, $\exists f \in \mathcal{F}in^c_+$ such that $f_{\mathcal{Q}^p_\phi}(p) > 0$ and for any process $\mathcal{S}$, $\forall s \in S$, $f_{\mathcal{S}}(s) > 0 \Rightarrow s \models \phi$. By Lemma 8.1, $f_{\mathcal{P}}(s) > f_{\mathcal{Q}^p_\phi}(s) > 0$, so f satisfies the conditions required by the lemma. □

Corollary 8.1 *Given $\phi \in \mathcal{L}_\vee$, a labelled Markov process $\mathcal{S}$, $c \in (0,1]$ and a state $s \in S$, if $s \models \phi$, then there exists $f \in \mathcal{F}in^c_+$ such that*

(1) $f_{\mathcal{S}}(s) > 0$;
(2) for any state s' of any other labelled Markov process $\mathcal{S}'$, we have $f_{\mathcal{S}'}(s') > 0 \Rightarrow s' \models \phi$.

Corollary 8.2 *Given $\phi \in \mathcal{L}_\vee$ and $c \in (0,1]$, there exists $f_\phi \in \mathcal{F}^c_+$ such that for every state s of any labelled Markov process $\mathcal{S}$,*

$$f_\phi(s) > 0 \Leftrightarrow\ s \models \phi.$$

Example 8.4 f_ϕ satisfies:

- $f_{\mathsf{T}} = \mathbf{1}$
- For any state s in process $\mathcal{S}$, $f_{\langle a \rangle_q \mathsf{T}}(s) = \max(\tau_a(s, S) - q, 0)$.

- $f_{\phi\wedge\psi} = \min(f_\phi, f_\psi)$
- $f_{\phi\vee\psi} = \max(f_\phi, f_\psi)$

The next result says that functions are sound and complete for bisimulation.

Theorem 8.1 *For any labelled Markov processes $\mathcal{S}$, $\mathcal{S}'$, $\forall c \in (0,1]$,*

$$s \in S \text{ and } s' \in S' \text{ are bisimilar iff } (\forall f \in \mathcal{F}in^c_+)\ [f_{\mathcal{S}}(s) = f_{\mathcal{S}}(s')]$$

Note that the left-to-right direction is also true for any functional of $\mathcal{F}^c$ but $\mathcal{F}in^c_+$ is enough for the other direction.

Proof. $(\Rightarrow)$: We show that for any bisimulation R, sRs' implies that $(\forall f \in \mathcal{F}^c)\ [f_{\mathcal{S}}(s) = f_{\mathcal{S}'}(s')]$. The proof proceeds by induction on the structure of the functional expression f. The key case is when f is of the form $(\langle a \rangle g)$. Then we would like to show that $\int_{t\in S} g(t)\tau_a(s,dt) = \int_{t\in S'} g(t)\tau'_a(s',dt)$. Consider any simple function h approximating g, with values $v_i, i = 1 \ldots n$, defined by $h(s) = max\{v_i \mid v_i \leq g(s)\}$. Then the set $S_i = h^{-1}(v_i) \subseteq S \cup S'$ is measurable because it is $g^{-1}([v_i, v_{i+1}))$ and it is R-closed because if $t \in S_i$ and tRt' then by induction $g(t) = g(t')$, so $t' \in S_i$. Thus $\tau_a(s, S_i) = \tau'_a(s', S_i)$, which shows the result.

$(\Leftarrow)$: Assume that s and s' are not bisimilar. Then there is a formula ϕ of $\mathcal{L}$ such that $s \models \phi$ and $s' \not\models \phi$ (or the converse). By Corollary 8.1, there is a functional expression $f \in \mathcal{F}in^c_+$ such that $f_{\mathcal{S}}(s) > 0$ and $f_{\mathcal{S}'}(s') = 0$. □

Given that we now know that functional expressions characterise bisimulation and that logical formulas also characterise bisimulation we immediately get:

Corollary 8.3 *For any process $\mathcal{S}$, $(\forall c \in (0,1])$, $\forall s, s' \in S$*

$$[(\forall\phi \in \mathcal{L})\ s \models_{\mathcal{S}} \phi \Leftrightarrow s' \models_{\mathcal{S}'} \phi] \Leftrightarrow (\forall f \in \mathcal{F}^c)\ [f_{\mathcal{S}}(s) = f_{\mathcal{S}'}(s')].$$

Note that for the $\mathcal{L}$ sub-fragment of the logic, the resulting function is in $\mathcal{F}in^c_+$.

The following example shows that the conditional functional expressions are necessary.

Example 8.5 Consider the processes A_1, A_2 in Figure 8.1. The calculations of Example 8.2 show that the s_0 states of A_1, A_2 are distinguishable.

Furthermore, the states are indistinguishable if we use only the functionals $\mathbf{1}, \mathbf{1} - f, \langle a \rangle f, \min(f_1, f_2), \sup_{i \in \mathbf{N}} f_i$. Thus, Example 8.2 shows that the functional expression $f \ominus q$ is indeed necessary.

So far we have shown that functional expressions are just as good for characterizing bisimulation as were logical formulas. We are now in a position to use the extra information in the functions to define a metric.

8.3 Metrics on Processes

In the present section we introduce the formal notion of pseudometrics between processes. As we have explained above, the metrics measure how different the processes are from the point of view of observable behaviour. We show that each $d^c, c \in (0,1]$ is a metric. In particular, processes at 0 distance are bisimilar.

Definition 8.3 Each collection $\mathcal{F}^c$ of functional expressions induces a distance function as follows:

$$d^c(\mathcal{P}, \mathcal{Q}) = \sup_{f \in \mathcal{F}^c} |f_{\mathcal{P}}(p_0) - f_{\mathcal{Q}}(q_0)|.$$

Theorem 8.2 *For all $c \in (0,1]$, d^c is a metric.*

Proof. The transitivity and symmetry of d^c are immediate. $d^c(\mathcal{S}, \mathcal{S}') = 0$ iff $\mathcal{S}$ and $\mathcal{S}'$ are bisimilar follows from Theorem 8.1. □

This definition is close in form to the definition of the Kantorovich metric [Hut81] which is used in the theory of optimal transport problems and also in the theory of fractals by Hutchinson. The difference is in the class of functions used. In the Kantorovich metric one uses the family of Lipschitz[1] functions. In our case the underlying state space is not a metric space[2] so we cannot really talk about Lipschitz functions. However – in a sense – these functions are really close to being Lipschitz. In suitable situations, one can show that our functions are dense in the class of Lipschitz functions.

We study the family of metrics $\{d^c \mid c \in (0,1]\}$. These metrics support the spectrum of possibilities of relative weighting of the two factors that contribute to the distance between processes: the complexity of the functions distinguishing them versus the amount by which each function distinguishes

[1] With Lipschitz constant 1 these are just the contractive functions.

[2] It is of course metrisable, being analytic.

them. d^1 captures only the differences in the probability numbers; probability differences at the first transition are treated on par with probability differences that arise very deep in the evolution of the process. In contrast, the metrics d^c for $c < 1$ give more weight to the probability differences that arise earlier in the evolution of the process, i.e. differences identified by simpler functions. As c approaches 0, the future gets discounted more.

As is usual with metrics, the actual numerical values of the metric are less important than the notions of convergence that they engender. Thus, we take the uniformity view of metrics, e.g. see [Ger85][3], and will view the metric via properties like the significance of zero distance, relative distance of processes, contractivity and the notion of convergence rather than a detailed justification of the exact numerical values.

Example 8.6 The analysis of Example 8.5 yields $d^c(A_1, A_2) = c^2/4$. This is witnessed by the functional $\langle a \rangle \min(\langle a \rangle \mathbf{1}, (\mathbf{1} - \langle a \rangle \mathbf{1}) \ominus (\mathbf{1} - c))$.

Example 8.7 Consider the family of processes $\{\mathcal{P}_\epsilon \mid 0 \leq \epsilon < r\}$ where $\mathcal{P}_\epsilon = a_{r-\epsilon}.\mathcal{Q}$, i.e. $\mathcal{P}_\epsilon$ is the process that does an a with probability $r - \epsilon$ and then behaves like $\mathcal{Q}$. The function expression $(\langle a \rangle \mathbf{1})$ evaluates to $(r - \epsilon)c$ at $\mathcal{P}_\epsilon$. This functional expression witnesses the distance between any two $\mathcal{P}$'s (other functions will give smaller distances). Thus, we get $d(\mathcal{P}_{\epsilon_1}, \mathcal{P}_{\epsilon_2}) = c|\epsilon_1 - \epsilon_2|$. This furthermore ensures that $\mathcal{P}_\epsilon$ converges to $\mathcal{P}_0$ as ϵ tends to 0.

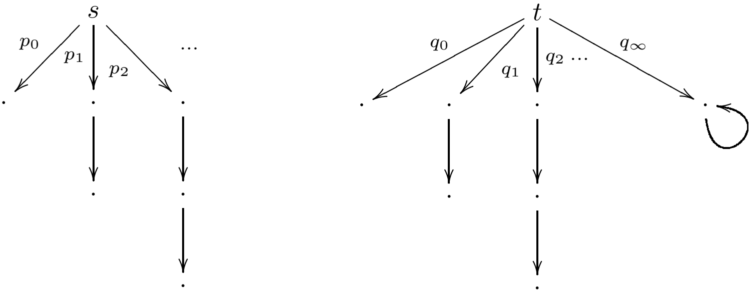


Fig. 8.3

[3] Intuitively, a uniformity captures relative distances, eg. is x is closer to z than y; it does not tell us what the actual distances are. For example, a uniformity on a metric space M is induced by the collection of all ϵ balls S_ϵ where $S_\epsilon = \{\{y \mid d(x, y) < \epsilon\} \mid x \in M\}$.

Example 8.8 (from [DEP98]) Consider processes s and t of Figure 8.3. t is just like s except that there is an additional transition to a state which then has an a-labelled transition back to itself. The probability numbers are as shown. If both processes have the same values on all functional expressions we will show that $q_\infty = 0$, i.e. it really cannot be present. The functional expression $(\langle a\rangle\mathbf{1})$ yields $c(\sum_{i\geq 0} p_i)$ on s and $c(q_\infty + \sum_{i\geq 0} q_i)$ on t. The functional expression $(\langle a\rangle\langle a\rangle\mathbf{1})$ yields $c^2(\sum_{i\geq 1} p_i)$ on s and $c^2(q_\infty + \sum_{i\geq 2} q_i)$ on t. Thus, we deduce that $p_0 = q_0$. Similarly, considering functional expressions $(\langle a\rangle\langle a\rangle\langle a\rangle\mathbf{1})$ etc, we deduce that $p_n = q_n$. Thus, $q_\infty = 0$.

8.4 Metric Reasoning for Process Algebras

In this section, we use a process algebra and an example coded in the process algebra to illustrate the type of reasoning provided by our study.

A process algebra

The process algebra we introduce describes probabilistically determinate processes. The processes are input-enabled [LT89; Dil88; Jos92] in a weak sense $((\forall p \in P)\ (\forall a \in \mathcal{A})\ \tau_{a?}(p, P) > 0)$ and communication is via CSP style broadcast. We could alternatively assume $(\forall s \in P)\ (\forall a \in \mathcal{A})\ \tau_{a?}(s, P) = 1$. All the results of this section continue to hold, the only change is in the definition of prefixing $a?_r.\mathcal{Q}$, where one adds a self-loop labelled $a?$, probability $1 - r$ to the start state.

The process combinators that we consider are parallel composition, prefixing and probabilistic choice. We do not consider hiding since this paper focuses on strong probabilistic bisimulation. Though we do not enforce the fact that output actions do not block, this assumption can safely be added to the process algebra[4].

We assume an underlying set of labels $\mathcal{A}$. Let $\mathcal{A}? = \{a? \mid a \in \mathcal{A}\}$ be the set of input labels, and $\mathcal{A}! = \{a! \mid a \in \mathcal{A}\}$ the set of output labels. Every process $\mathcal{P}$ is associated with a subset of labels: $\mathcal{P}_O \subseteq \mathcal{A}!$, the set of relevant output labels. This signature is used to constrain parallel composition.

[4] This would make it an IO calculus [Vaa91].

Prefixing
$\mathcal{P} = a?_r.\mathcal{Q}$, where r is a rational number, is the process that accepts input a and then performs as $\mathcal{Q}$. The number r is the probability of accepting $a?$. With probability $(1 - r)$ the process $\mathcal{P} = a?_r.\mathcal{Q}$ will *block* on an $a?$ label. P is given by adding a new initial state, p_0 to Q. Add a transition labelled $a?$ from p_0 to the start state of $\mathcal{Q}$ with probability r. For all other labels l, add a $l?$ labelled self-loop at p_0 with probability 1.

Output prefixing, $\mathcal{P} = a!_r.\mathcal{Q}$, where r is a rational number, the process that performs output action $a!$ and then functions as $\mathcal{Q}$, is defined analogously. In this case, $\mathcal{P}_O = \mathcal{Q}_O \cup \{a!\}$. For both input and output prefixing, we have: $d^c(a_r.\mathcal{P}, a_u.\mathcal{P}) \leq c \mid r - u \mid$.

Probabilistic Choice
$\mathcal{P} = \mathcal{Q} +_r \mathcal{Q}'$ is the probabilistic choice combinator that chooses $\mathcal{Q}$ with probability r and $\mathcal{Q}'$ with probability $1 - r$. $\mathcal{P}_O = \mathcal{Q}_O \cup \mathcal{Q}'_O$. $P = Q \uplus Q'$. Now $\tau_a^{\mathcal{P}}(q, X \uplus X') = \tau_a(q, X)$ if $q \in Q$, and $\tau_a^{\mathcal{P}}(q, X \uplus X') = \tau'_a(q, X')$ if $q \in Q'$. We define an initial distribution μ: $\mu(\{q_0\}) = r, \mu(\{q'_0\}) = 1 - r$, it is clear that this could be defined in the initial state format (see [DGJP04]).

We have: $d^c(\mathcal{P} +_r \mathcal{Q}, \mathcal{P} +_u \mathcal{Q}) \leq \mid r - u \mid d^c(\mathcal{P}, \mathcal{Q})$; $d^c(\mathcal{P} +_r \mathcal{Q}, \mathcal{P}' +_r \mathcal{Q}) \leq r d^c(\mathcal{P}, \mathcal{P}')$.

Parallel Composition
$\mathcal{P} = \mathcal{Q} \mid\mid \mathcal{Q}'$ is permitted if the output actions of $\mathcal{Q}, \mathcal{Q}'$ are disjoint, i.e. $\mathcal{Q}_O \cap \mathcal{Q}'_O = \emptyset$. The parallel composition synchronises on all labels in $\mathcal{Q}_L \cap \mathcal{Q}'_L$. $\mathcal{P}_O = \mathcal{Q}_O \uplus \mathcal{Q}'_O$. $P = Q \times Q'$. The definition of $\tau_a^{\mathcal{P}}$ is motivated by the following idea. Let s (resp. s') be a state of $\mathcal{Q}$ (resp. $\mathcal{Q}'$). We expect the following synchronised transitions from the product state (s, s'). The disjointness of the output labels of $\mathcal{Q}, \mathcal{Q}'$ ensures that there is no non-determinism. Formally, if $l = a! \in \mathcal{Q}_O$, then $\tau_{a?}^{\mathcal{P}}((s, s'), (t, t')) = \tau_{a!}^{\mathcal{P}}((s, s'), (t, t')) = \tau_{a!}(s, t) \times \tau'_{a?}(s', t')$. The case when $a! \in \mathcal{Q}'_O$ and $l = a?$ is similar.

To fix terminology, let us use the same symbol $\mathcal{P}$ to stand for the syntactic expression for a process and for the labelled transition system generated by $\mathcal{P}$. When a process, say $\mathcal{P}$, has an a-transition we cannot say that it results in a process $\mathcal{P}'$; instead, we must say that it results in some distribution of possible states of $\mathcal{P}$ – these states are, of course, denoted in the syntax by derivatives of the syntactic expression for $\mathcal{P}$.

Definition 8.4 Let $\mathcal{P}$ be a process. Then $\mathcal{P}$ **after** a is the same process but with start distribution given by $\nu(t) = \tau_a(p_0, t)$. We perform some normalisation based on the total probability of the resulting initial configuration $\nu(P)$: if $\nu(P) > 0$, it is normalised to be 1; if $\nu(P) = 0$, it is left untouched. This definition extends inductively to $\mathcal{P}$ **after** α, where α is a finite sequence of labels $(a_0, a_1, a_2, \ldots, a_k)$.

Note that $\mathcal{P}$ **after** α is identical to $\mathcal{P}$ – i.e. it denotes the same labelled transition system – except that its initial configuration may be different.

Lemma 8.4 *Let $h \in \mathcal{F}^c$, let $\mathcal{P}$ be a process and let $a \in \mathcal{A}$. Then*

$$\langle a \rangle h(p_0) = c \times h(\mathcal{P} \textbf{ after } a).$$

*Here $h(\mathcal{P}$ **after** $a)$ means $h(p_0')$ where p_0' is the initial state of $\mathcal{P}$ **after** a.*

Theorem 8.3 ***Contractivity of process combinators***

- *$d(l_r.\mathcal{P}, l_r.\mathcal{Q}) \leq cd(\mathcal{P}, \mathcal{Q})$ for any label l*
- *$d(\mathcal{P} +_r \mathcal{R}, \mathcal{Q} +_r \mathcal{R}) \leq d(\mathcal{P}, \mathcal{Q})$ for any $\mathcal{R}$*
- *$d(\mathcal{P} \mid\mid \mathcal{R}, \mathcal{Q} \mid\mid \mathcal{R}) \leq d(\mathcal{P}, \mathcal{Q})$ for any $\mathcal{R}$ for which $\mathcal{P} \mid\mid \mathcal{R}, \mathcal{Q} \mid\mid \mathcal{R}$ are defined.*

Proof. The proof proceeds by induction on functional expressions. Let $f^-(\mathcal{P}, \mathcal{Q})$ mean $|f(p_0) - f(q_0)|$ where p_0 (q_0) is the initial state of $\mathcal{P}$ $(\mathcal{Q})$. We show that for any f, there exists a g such that f^- of the lhs is less than or equal to some g^- of the rhs. We omit the detailed calculations and prove the result for the key case where f is $\langle a \rangle h$, for parallel composition. Let $\mathcal{P}' = \mathcal{P}$ **after** $b?$, $\mathcal{Q}' = \mathcal{Q}$ **after** $b?$ and $\mathcal{R}' = \mathcal{R}$ **after** $b!$. By induction, we know that there is some functional g such that $h^-(\mathcal{P}' \mid\mid \mathcal{R}', \mathcal{Q}' \mid\mid \mathcal{R}') \leq g^-(\mathcal{P}', \mathcal{Q}')$. Now suppose $a = b!$, and $b! \in \mathcal{R}_O$, then $\mathcal{P} \mid\mid \mathcal{R}$ **after** $a = \mathcal{P}' \mid\mid \mathcal{R}'$. Now we calculate as follows:

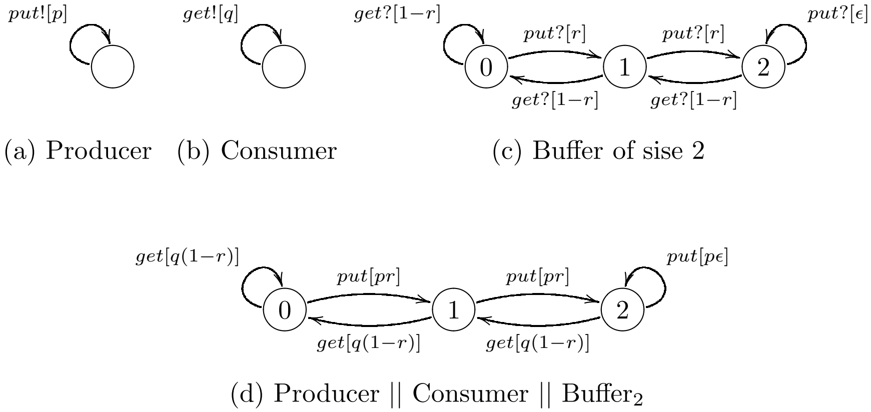
$$\begin{aligned}
(\langle a \rangle h)^-(\mathcal{P} \mid\mid \mathcal{R}, \mathcal{Q} \mid\mid \mathcal{R}) &= c \times h^-((\mathcal{P} \mid\mid \mathcal{R}) \textbf{ after } a, (\mathcal{Q} \mid\mid \mathcal{R}) \textbf{ after } a) \\
&= c \times h^-(\mathcal{P}' \mid\mid \mathcal{R}', \mathcal{Q}' \mid\mid \mathcal{R}') \\
&\leq c \times g^-(\mathcal{P}', \mathcal{Q}') \\
&= (\langle a \rangle g)^-(\mathcal{P}, \mathcal{Q}).
\end{aligned}$$

□

Thus, Theorem 8.1 allows us to conclude that bisimulation is a congruence with respect to these operations.

A bounded buffer example

We specify a producer consumer process with a bounded buffer (along the lines of [PS85]). The producer is specified by the one-state finite automaton shown in Figure 8.4(a) – it outputs a *put*, corresponding to producing



(a) Producer (b) Consumer (c) Buffer of sise 2

(d) Producer || Consumer || Buffer$_2$

Fig. 8.4 The producer consumer example

a packet, with probability p. To keep the figure uncluttered, we omit the input-enabling arcs, all of which have probability 1. The consumer (Figure 8.4(b)) is analogous – it outputs a *get* with probability q, corresponding to consuming a packet. The buffer is an n-state automaton, the states are merely used to count the number of packets in the buffer, while the probabilities code up the probability of scheduling either the producer or the consumer (thus the producer gets scheduled with probability r, and then produces a packet with probability p). Upon receiving a *put* in the last state, the buffer accepts it with a very small probability ϵ, modelling a blocked input. The parallel composition of the three processes is shown in Figure 8.4(d). Notice that the behavior of this process is very similar to a random walk – the process moves to the next state with probability $r = p(1-q)/(p+q-pq)$, corresponding to a *put*, and the previous state with probability $1-r$, corresponding to a *get* – we ignore the transitions back to the same state, regarding them as no-ops. It is easy to show that in any run of this process with a large number of *put* actions, the expected fraction of discarded packets is approximately $(1 - r/r)^{-n}$ – we compute the stationary distribution for this process, and since it is ergodic, this stationary distribution is reached after a large number of steps. Then the *put* actions in the last state result in lost packets.

As the buffer size increases, the distance between the bounded buffer and the unbounded buffer decreases to 0. Let $\mathcal{P}_k =$ Producer $||$ Consumer $||$ Buffer$_k$, where Buffer$_k$ denotes the process Buffer with k states. Then by looking at the structure of the process, we can compute that $d(\mathcal{P}_k, \mathcal{P}_\infty) \propto (cpr)^k$. Thus we conclude the following:

- As the bounded buffer becomes larger, it approximates an infinite buffer more closely: if $m > k$ then $d^c(\mathcal{P}_k, \mathcal{P}_\infty) > d^c(\mathcal{P}_m, \mathcal{P}_\infty)$.
- As the probability of a put decreases, the bounded buffer approximates an infinite buffer more closely. Thus if $p < p'$, $d^c(\mathcal{P}^p, \mathcal{P}^p_\infty) < d^c(\mathcal{P}^{p'}, \mathcal{P}^{p'}_\infty)$, where the superscripts indicate the producer probability.
- Similarly, as the probability of scheduling the Producer process (r) decreases, the buffer approximates an infinite buffer more closely.

8.5 Perturbation

One of the major criticisms of process equivalences is that they are not robust. The results of this section show that if one slightly perturbs the probabilities in a process the result is close.

Definition 8.2 Let $\mathcal{S} = (S, s_0, \Sigma, \tau)$ be a labelled Markov process. Define $\mathcal{S}' = (S, s_0, \Sigma, \tau')$ to be an ϵ-perturbation of $\mathcal{S}$ if for all labels a,

$$\forall s \in S.\ \forall X \in \Sigma.\ |\tau_a(s, X) - \tau'_a(s, X)| < \epsilon.$$

Our metric accommodates the notion of small perturbations of probabilities.

Proposition 8.3 *If $c < 1$, and $\mathcal{S}'$ is an ϵ-perturbation of $\mathcal{S}$, then $d^c(\mathcal{S}, \mathcal{S}') < k\epsilon$ where $k = \sup_n nc^n$.*[5]

Proof. The proof is by induction on the formulas. The sole non-trivial case is $\langle a \rangle f$. We write f for $f_{\mathcal{S}}$ and f' for $f_{\mathcal{S}'}$. Let $depth(f) = n$, and $|f(t) - f'(t)| < \epsilon nc^n$. Then $f(s) \leq c^n$ and

[5] *e.g.* $k = 1$ for $c \leq 1/2$.

$$\begin{aligned}
c[\int_t f(t)\tau_a(s,dt) - \int_t f'(t)\tau'_a(s,dt)] \\
&= c\int f(t)[\tau_a(s,dt) - \tau'_a(s,dt)] + c\int \tau'_a(s,dt)[f(t) - f'(t)] \\
&< c^{n+1}|\tau(s,X) - \tau'(s,X)| + nc^{n+1}\epsilon\int \tau'_a(s,t) \\
&< c^{n+1}\epsilon + nc^{n+1}\epsilon \\
&= (n+1)c^{n+1}\epsilon.
\end{aligned}$$

Here X is the set on which the measure $\tau_a(s,.) - \tau'_a(s,.)$ is positive. □

For $c = 1$, nc^n increases without limit, and Example 8.3 shows that the above lemma does not hold for $c = 1$. However in this case we can still perturb the process $\mathcal{S}$ in the following way – let $\mathcal{S}$ be unfolded, so it has no loops. Let $\epsilon_i, i \in \mathbf{N}$ be nonnegative rationals such that $\sum_i \epsilon_i = \epsilon < 1/3$. Now we obtain $\mathcal{S}'$ by taking the same state set as $\mathcal{S}$, and for each state s at depth n, $|\tau_a(s,X') - \tau'_a(s,X')| < \epsilon_n$ for each label a and each measurable set X'. Then we can show by a similar calculation as above that $d^1(\mathcal{S},\mathcal{S}') < 1 - e^{-2\epsilon}$, thus as $\epsilon \longrightarrow 0$, $d^1(\mathcal{S},\mathcal{S}') \longrightarrow 0$.

Example 8.9 Consider "straight line" formulas generated by

$$\phi ::= \mathsf{T} \mid \langle a\rangle_q \phi$$

Consider one such $\phi = \langle a_1\rangle_{q_1} \ldots \langle a_n\rangle_{q_n}\mathsf{T}$. Let $\mathcal{P}$ be a finite-state process unfolded to the depth of the formula such that p_0, the start state of $\mathcal{P}$, satisfies the formula. An easy induction, using the proof of Lemma 8.2, shows that

$$f_\phi(p_0) \geq c^n \prod_i (r_i - q_i)$$

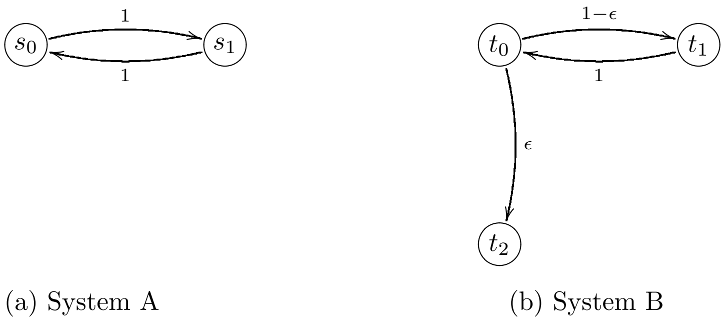
where $r_i = \inf_{s\in X_i} \tau_{a_i}(s, X_{i+1})$ and X_{i+1} is the set of all the states in level $i+1$ which satisfy the suffix formula $\langle a_{i+1}\rangle_{q_{i+1}} \ldots \langle a_n\rangle_{q_n}\mathsf{T}$. Note that this bound is achieved by the n-length chain automaton which has transition probabilities r_i.

The form of the expression $f(p_0) \geq c^n \prod_i (r_i - q_i)$ tells us that if $f(p_0) > \epsilon$, we can perturb the probabilities at some level by up to $\epsilon^{\frac{1}{n}}/c$, and the resulting process will continue to satisfy the formula.

Finally we close with an important example that shows the importance of the connectivity of the transition graph.

Example 8.10 Consider the systems shown in Fig. 8.5. The states s_0 and t_0 appear to be very similar and are clearly metrically close - or are they? In system A there is no steady state distribution (the Markov chain fails to be aperiodic) whereas in system B there is a steady state, namely all the mass eventually leaks into state t_2 and stays there. How is it that the asymptotic behaviour can be so drastically different when the states are so close?

The short answer is that the states *are not at all close.* If one computes the distance, a routine calculation shows that the states s_0 and t_0 are at distance 1 for the metrics with $c = 1$ – the maximum possible distance! Even with $c < 1$ the distance is large though less than 1.



(a) System A

(b) System B

Fig. 8.5 The effect of topology change

8.6 The Asymptotic Metric

Often in probabilistic reasoning one is interested in the behaviour of a process once it has “settled down” into some steady state. For LMPs there may not be a steady state but there is a notion asymptotic metric that can be defined for our simple process algebra.

Define the j distance between $\mathcal{P}, \mathcal{Q}$,

$$d^c_j(\mathcal{P}, \mathcal{Q}) = \sup\{d^c(\mathcal{P}\ \mathbf{after}\ \alpha, \mathcal{Q}\ \mathbf{after}\ \alpha) \mid \mathit{length}(\alpha) = j\}.$$

We define the asymptotic distance between processes $\mathcal{P}$ and $\mathcal{Q}$, $d^c_\infty(\mathcal{P}, \mathcal{Q})$

to be

$$d^c_\infty(\mathcal{P}, \mathcal{Q}) = \limsup_{i\to\infty}\, \sup_{j>i} d^c_j(\mathcal{P}, \mathcal{Q}).$$

The fact that d^c_∞ satisfies the triangle inequality and is symmetric immediately follows from the same properties for d.

Example 8.11 For any process $\mathcal{P}$, $d^c_\infty(a_q.\mathcal{P}, a_r.\mathcal{P}) = 0$, where $q, r > 0$. Consider A_3 from Figure 8.2. Without the normalisation in the definition of A_3 **after** α, we would have got $d^c_\infty(a_q.A_3, a_r.A_3) = c|q - r|$

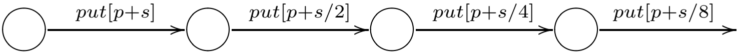


Fig. 8.6 A producer with transient behavior

Example 8.12 Consider the producer process $\mathcal{P}_2$ shown in Figure 8.6. This is similar to the producer $\mathcal{P}_1$ in Figure 8.4, except that initially the probability of producing *put* is more than q, however as more *put*s are produced, it asymptotically approaches q. If we consider the asymptotic distance between these two producers, we see that $d^c(\mathcal{P}_2 \textbf{ after } put^n, \mathcal{P}_1 \textbf{ after } put^n) \propto 2^{-(n+1)}$. Thus $d^c_\infty(\mathcal{P}_1, \mathcal{P}_2) = 0$. Now by using the compositionality of parallel composition, we see that $d^c_\infty(\mathcal{P}_1 \mid\mid \text{Consumer} \mid\mid \text{Buffer}_k, \mathcal{P}_2 \mid\mid \text{Consumer} \mid\mid \text{Buffer}_k) = 0$, which is the intuitive result.

Asymptotic equivalence is preserved by parallel composition and prefixing.

Theorem 8.4

(1) $d^c_\infty(l_r.\mathcal{P}, l_r.\mathcal{Q}) \leq d^c_\infty(\mathcal{P}, \mathcal{Q})$ *for any label* l.
(2) $d^c_\infty(\mathcal{P} \mid\mid \mathcal{R}, \mathcal{Q} \mid\mid \mathcal{R}) \leq d_\infty(\mathcal{P}, \mathcal{Q})$.

For the key case of parallel composition, the proof is based on:

$$(\mathcal{P} \mid\mid \mathcal{Q}) \textbf{ after } \alpha = (\mathcal{P} \textbf{ after } \alpha_1) \mid\mid (\mathcal{Q} \textbf{ after } \alpha_2),$$

where α_1 has the $a!$ labels of α replaced by $a?$ where $a! \notin \mathcal{P}_O$, and similarly for α_2.

8.7 Behavioural Properties of the Metric

The theme of this section is extracting behavioural information from the metric. Our first lemma shows that we can find a function – in our class of functions – which is characteristic for ϵ-balls around a given state. Many of the proofs are omitted, they appear in Desharnais's thesis [Des99].

Lemma 8.5 *Let s be a state in a labelled Markov process $\mathcal{S}$, $\epsilon \in (0, 0.5)$ and $c \in (0, 1]$. Let $\{f_i \mid i = 1 \ldots n\}$ be a finite set of functional expressions of $\mathcal{F}^c$. Then, there is a single function $f \in \mathcal{F}^c$ such that $f(s) = \epsilon$, and for every $t \in S$ we have $f(t) = 0$ iff for any i,$|f_i(s) - f_i(t)| \geq \epsilon$.*

Proof. Define functional expressions g_i as follows. Let $f_i(s) = q_i$.

$$g_i = \begin{cases} \min(f_i \ominus (q_i - \epsilon), (\mathbf{1} - f_i) \ominus (1 - q_i - \epsilon)), \text{if } q_i \geq \epsilon \\ (\mathbf{1} - f_i) \ominus (1 - q_i - \epsilon), \text{if } q_i < \epsilon \end{cases}$$

Note that $g_i(s) = \epsilon$. Also, for any state t in any process, if $f_i(t) \geq q_i + \epsilon$ or $f_i(t) \leq q_i - \epsilon$, then $g_i(t) = 0$. The functional expression $f = \min(g_1, \ldots, g_n)$ satisfies the required properties. □

8.8 The Pseudometric as a Maximum Fixed Point

So far we have defined the metric as arising from a real-valued modal logic. This is very useful if one wants to show that two processes are at a distance of *at least* say α: one just has to find a function that differs by α on the two processes. On the other hand if one wants to perform coinductive style reasoning – as is familiar with bisimulation in process algebras – one needs a fixed-point definition of the metric. We present just such a definition; it will be very much like the fixed point definition of bisimulation. This idea is due to van Breugel and Worrell who gave a category-theoretic presentation and defined a final coalgebra in a category of metric spaces. We are essentially using the same idea except that we work with a lattice of metrics and use order-theoretic fixed-point theory rather than category theory.

Throughout this section we work with a finite-state system rather than a full blown LMP. This allows a much simpler presentation and, more importantly, the use of techniques from linear programming. Everything can be done in greater generality using the more powerful tools of infinite-dimensional linear programming but this would involve a long mathematical digression.

We work with labelled Markov chains, i.e. the discrete version of an LMP. We recapitulate the definition below.

Definition 8.5 A labelled Markov chain (henceforth LMC), is a tuple $\mathcal{S} = (S, \mathcal{A}, \forall a \in \mathcal{A}\tau_a, s_0)$, where
(1) S is a countable set of states, s_0 is the start state.
(2) $\mathcal{A}$ is a finite set of action symbols,
(3) For each $a \in \mathcal{A}$ the transition function $\tau_a : S \times S \longrightarrow [0,1]$ is a subprobability distribution on its second argument.

We fix an LMC and consider pseudometrics on its set of states. We work with pseudometrics where the maximum distance is 1[6]; so-called *one-bounded* pseudometrics.

Definition 8.6 $\mathcal{M}$ is the class of 1-bounded pseudometrics on states with the ordering

$$m_1 \preceq m_2 \text{ if } (\forall s,t)\ [m_1(s,t) \geq m_2(s,t)]$$

Lemma 8.6 $(\mathcal{M}, \preceq)$ *is a complete lattice.*

Proof. The least element is given by: $\perp(s,t) = 0$ if $s = t, 1$ otherwise. The top element is given by $(\forall s,t)\top(s,t) = 0$. Greatest lower bounds are given by: $(\sqcap\{m_i\})(s,t) = \sup_i m_i(s,t)$ Note that:

$$\begin{aligned}(\sqcap\{m_i\}(s,t) &= \sup_i m_i(s,t)\\ &\leq \sup_i [m_i(s,u) + m_i(t,u)]\\ &\leq \sup_i [m_i(s,u)] + \sup_i [m_i(t,u)]\\ &\leq (\sqcap\{m_i\}(s,u) + (\sqcap\{m_i\}(u,t)\end{aligned}$$

□

$m \in \mathcal{M}$ is extended to distributions on sets of states as follows. The definition is based on what is variously called the Kantorovich metric, the Vaserstein (or Wasserstein) metric and the Hutchinson metric on probability measures; we have merely simplified the definition for our context of discrete finite state distributions. We will refer to it as the Kantorovich metric.

Definition 8.7 Let $m \in \mathcal{M}$. Let P, Q be distributions on states such that the total mass of P is not less than the total mass of Q. Then $m(P,Q)$

[6] Given a metric d one can always define a new metric $d'(x,y) = \frac{d(x,y)}{1+d(x,y)}$ with the same topology; so no real generality is lost.

is given by the solution to the following linear program:

$$\begin{aligned} \max \sum_i (P(s_i) - Q(s_i))a_i \\ \text{subject to : } \forall i.0 \leq a_i \leq 1 \\ \forall i,j.\ a_i - a_j \leq m(s_i, s_j). \end{aligned}$$

We need the constraints $a_i \leq 1$ if the distributions do not have equal total probability, without them the maximum is unbounded.

One of the most useful and fundamental ideas in linear programming is the *duality* principle [Chv83]. According to this a linear program expressed as a maximisation principle is equivalent to a dual linear program expressed as a minimisation principle, and vice versa. We shall use this idea repeatedly in what follows. Why is this useful? The original linear program can be used easily to provide lower bounds on the solution whereas the dual can be used to provide upper bounds. Thus one can often prove equations involving the solutions of an LP by computing matching upper and lower bounds. This duality principle can be extended to certain infinite dimensional cases but one has to take care that appropriate topological conditions are met [AN87].

Van Breugel and Worrell showed how to view the definition of the Kantorovich metric as a transportation problem. The idea of the dual is that one has to "ship" the probability mass from one place to another in order to transform P to Q with the metric on states giving a cost for moving the probability mass; the dual is then asking for the minimum cost transformation of P into Q. Following the analysis of [vBW01], $m(P,Q)$ is given by the solution to the following dual linear program:

$$\begin{aligned} \min \sum_{i,j} l_{ij} m(s_i, s_j) + \sum_i x_i + \sum_j y_j \\ \text{subject to : } \forall i. \textstyle\sum_j l_{ij} + x_i = P(s_i) \\ \forall j. \textstyle\sum_i l_{ij} + y_j = Q(s_j) \\ \forall i,j.\ l_{ij}, x_i, y_j \geq 0. \end{aligned}$$

The following lemma shows that this extension to distributions satisfies the triangle inequality and is consistent with the ordering on pseudometrics. The proof of the first item is an elementary exercise using the primal linear program. The proof of the second item uses the dual linear program – every solution to the (dual) linear program $m'(P,Q)$ is also a solution to the (dual) linear program for $m(P,Q)$.

Lemma 8.7

- *Let $m \in \mathcal{M}$. Then, $(\forall P, Q, R)$ $m(P,Q) \leq m(P,R) + m(Q,R)$.*
- *Let $m, m' \in \mathcal{M}$ such that $m \preceq m'$. Then, for all distributions on states P, Q, $m(P,Q) \geq m'(P,Q)$.*

Proof.

- For the first item, we proceed as follows. We first prove that if total mass of P is less than the total mass of Q, then

  $$\max \sum_i (P(s_i) - Q(s_i))a_i < \max \sum_i (Q(s_i) - P(s_i))a_i$$

  where the a_i are subject to the usual constraints $0 \leq a_i \leq 1, a_i - a_j \leq m(s_i, s_j)$.

  $$\begin{aligned}\sum_i (P(s_i) - Q(s_i))a_i &= \sum_i (Q(s_i) - P(s_i))(1 - a_i) - \sum_i Q(s_i) + \sum_i P(s_i) \\ &< \sum_i (Q(s_i) - P(s_i))(1 - a_i)\end{aligned}$$

  Now all the $b_i = 1 - a_i$ also satisfy the constraints on the a_i above, proving the result.
  To prove the triangle inequality, given distributions P, Q, R, $\sum_i (P(s_i) - Q(s_i))a_i = \sum_i (P(s_i) - R(s_i))a_i + \sum_i (R(s_i) - Q(s_i))a_i$. Taking the maximum over the a_i for the left side, we get $m(P,Q) \leq m(P,R) + m(R,Q)$.
- For the second item, note that every solution to the linear program defining $m'(P,Q)$ is also a solution to the linear program defining $m(P,Q)$. So, the maximum value $m(P,Q)$ is $\geq m'(P,Q)$.

□

The dual linear program is a key tool to move from distributions to states in the following sense. Given close-by distributions P, Q, the dual linear program permits us to construct a matching of states (that may include "splitting" of the probabilities assigned by P, Q), in such a way that exactly the distance between P, Q can be recovered.

Lemma 8.8 *Let P and Q be probability distributions on a set of states K. Let P_1 and P_2 be such that: $P = P_1 + P_2$. Then, there exist Q_1, Q_2, such that $Q_1 + Q_2 = Q$ and*

$$m(P,Q) = m(P_1, Q_1) + m(P_2, Q_2).$$

Proof. Let $\{l_{ij}\}$, $\{x_i\}$ and $\{y_j\}$ be such that the minimum is attained in the dual linear program above: that is, $m(P,Q) = \sum_{i,j} l_{ij} m(s_i,s_j) + \sum_i x_i + \sum_j y_j$. Define: $Q_k(s_j) = \sum_i l_{ij} P_k(s_i)/P(s_i) + y_j P_k(K)/P(K)$, for $k=1,2$. Then, $Q_1+Q_2=Q$.

Furthermore, setting

$l^1_{ij} = [P_1(s_i)/P(s_i)]l_{ij}$,
$y^1_j = [P_1(K)/P(K)]y_j$ and
$x^1_i = [P_1(s_i)/P(s_i)]x_i$

we get:

$$\begin{aligned}\sum_i l^1_{ij} + y^1_j &= Q_1(s_j)\\ \sum_j l^1_{ij} + x^1_i &= \sum_j [P_1(s_i)/P(s_i)]l_{ij} + [P_1(s_i)/P(s_i)]x_i\\ &= [P_1(s_i)/P(s_i)](\sum_j l_{ij} + x_i)\\ &= [P_1(s_i)/P(s_i)]P(s_i)\\ &= P_1(s_i).\end{aligned}$$

Thus,

$$m(P_1,Q_1) \le \sum_{i,j} l^1_{ij} m(s_i,s_j).$$

Similarly,

$$m(P_2,Q_2) \le \sum_{i,j} l^2_{ij} m(s_i,s_j).$$

Thus:

$$\begin{aligned}m(P_1,Q_1)+m(P_2,Q_2) &\le \sum_{i,j} l^1_{ij} m(s_i,s_j) + \sum_i x^1_i + \sum_j y^1_j\\ &\quad + \sum_{i,j} l^2_{ij} m(s_i,s_j) + \sum_i x^2_i + \sum_j y^2_j\\ &= \sum_{i,j}[l^1_{ij}+l^2_{ij}]m(s_i,s_j) + \sum_i (x^1_i+x^2_i) + \sum_j (y^1_j+y^2_j)\\ &= \sum_{i,j} l_{ij} * m(s_i,s_j) + \sum_i x_i + \sum_j y_j\\ &= m(P,Q).\end{aligned}$$

To show that $m(P_1,Q_1)+m(P_2,Q_2) \ge m(P,Q)$, consider the a_i that achieves the maximum in the definition of $m(P,Q)$. Recall that if $P(K) \ge$

$Q(K)$, then $P_i(K) \geq Q_i(K)$. Thus

$$\begin{aligned} m(P_1, Q_1) + m(P_2, Q_2) &\geq \sum_i (P_1(s_i) - Q_1(s_i))a_i + \sum_i (P_2(s_i) - Q_2(s_i))a_i \\ &= \sum_i (P(s_i) - Q(s_i))a_i \\ &= m(P, Q). \end{aligned}$$

□

As a straightforward corollary, we get a complete matching on individual states.

Corollary 8.4 *Given distributions P, Q there exist distributions P_i, Q_i such that*

- *P_i are point distributions that are non-zero at only one state.*
- *$P = \sum P_i$, $Q = \sum Q_i$*
- *$m(P, Q) = \sum m(P_i, Q_i)$.*

We now define a functional F on $\mathcal{M}$ that closely resembles the usual functional for bisimulation.

Definition 8.8 Define F, a functional on $\mathcal{M}$ as follows. $F(m)(s,t) < \epsilon$ if:

- $(\forall s \xrightarrow{a} P)$ $(\exists t \xrightarrow{a} Q)$ $[m(P, Q) < \epsilon]$.
- $(\forall t \xrightarrow{a} Q)$ $(s \xrightarrow{a} P)$ $[m(P, Q) < \epsilon]$.

$F(m)$ is well-defined because of the following lemma. The triangle inequality on $F(m)$ follows from the triangle inequality on m extended to distributions.

Lemma 8.9 *$F(m)$ is a pseudometric given by:*

$$F(m)(s,t) = \max(\max_{a \in \mathcal{A}} \sup_{s \xrightarrow{a} P} \inf_{t \xrightarrow{a} Q} m(P, Q), \max_{a \in \mathcal{A}} \sup_{t \xrightarrow{a} Q} \inf_{s \xrightarrow{a} P} m(P, Q)).$$

Proof. We prove the triangle inequality. Let $F(m)(s,t) < \epsilon_1, F(m)(t,u) < \epsilon_2$. Let $s \xrightarrow{a} P$. Since $F(m)(s,t) < \epsilon_1$, there exists a $t \xrightarrow{a} Q$ such that $m(P, Q) < \epsilon_1$. Since $F(m)(t,u) < \epsilon_2$, there exists a $u \xrightarrow{a} R$ such that $m(Q, R) < \epsilon_2$. From the triangle inequality on m (extended to distributions), $m(P, R) < \epsilon_1 + \epsilon_2$. □

Lemma 8.10 *F is monotone on $\mathcal{M}$.*

Proof. Let $m_2 \preceq m_1$. We need to show that $F(m_2) \preceq F(m_1)$, i.e. $(\forall s,t)F(m_1)(s,t) \leq F(m_2)(s,t)$.

Let $F(m_2)(s,t) < \epsilon$. Then,

- For all transitions $s \xrightarrow{a} P$, there exists $t \xrightarrow{a} Q$ such that $m_2(P,Q) < \epsilon$.
- For all transitions $t \xrightarrow{a} Q$, there exists a transition $s \xrightarrow{a} P$ such that $m_2(P,Q) < \epsilon$.

Since, $m_2(P,Q) \geq m_1(P,Q)$, it follows that $F(m_1)(s,t) < \epsilon$ as required. □

Using the basic fixed point theorem for complete lattices, F has a maximum fixed point.

The maximal fixed point of F is sound with respect to bisimulation. The forward implication of the proof uses the pseudometric m' defined as: $m'(s,t) = 0$ iff s and t are bisimilar and 1 otherwise. m' satisfies $m' \preceq F(m')$. The converse proceeds by showing that the equivalence relation R induced by 0 distance is a bisimulation.

Lemma 8.11 $s \approx t \Leftrightarrow m(s,t) = 0$*, where* m *is the maximum fixed point of* F*.*

Proof. For the forward direction, consider the pseudometric m' defined as: $m'(s,t) = 0$ iff s and t are bisimilar and 1 otherwise. Clearly $F(m') \leq m'$.

For the converse, consider the relation R induced by 0 metric distance. Clearly, this is an equivalence relation. We show that this equivalence relation is a bisimulation. Let $m(s,t) = 0$. Suppose that $s \xrightarrow{a} P$ and $P(E) = p$ for some R-closed set E. Given any $\epsilon > 0$, since $m(s,t) = 0$, we get a transition $t \xrightarrow{a} Q$ such that $m(P,Q) < d\epsilon/n^2$, where n is the number of states and $d = \min\{m(s_i,s_j) | s_i, s_j \text{ are states in the system}, m(s_i,s_j) > 0\}$. Then in the dual linear program, $l_{ij} < \epsilon/n^2$ for all i,j such that $m(s_i,s_j) > 0$, and so are x_i and y_j. Now $|P(s_i) - Q(s_i)| = \sum_j l_{ij} + x_i - \sum_k l_{ki} - y_i < \epsilon/n$ as l_{ii} cancels out, and there are at most n positive and n negative terms on the rhs. Thus $|P(E) - Q(E)| < \epsilon$, as desired. □

This fixed point definition of the metric can be shown to be the same as the previous definition. In a paper on weak bisimulation by Desharnais et. al. [DGJP02a] the metric analogue of weak bisimulation is developed in both ways and their equivalence is established. The LP techniques are crucial to this proof.

Chapter 9

Approximating Labelled Markov Processes

In the previous chapters we developed a theory of labelled Markov processes on continuous state spaces. How does one deal with these spaces computationally? Can one make available the model checking techniques that have been developed for finite-state probabilistic systems [BdA95; Bai96; HK97; BCHG^{+}97; dAKN^{+}00]? Clearly one needs to be able to approximate labelled Markov processes, but we need to do so in a way that respects the *behavioural* properties that we have taken such pains to analyse and characterise.

Typically one approximates a continuous-state system by "carving up" the state space into a finite number of clusters. These clusters are usually based on some "natural" geometric structure of the state space. Quite often this works very well, but in general, the geometry of the state space is not guaranteed to be a good guide to its dynamics. What we need is an approximation scheme that is guided by the bisimulation relation. This is precisely what we do in this chapter.

We show that, for every **LMP** $\mathcal{S}$, we can define a sequence of *finite-state* approximants $\mathcal{S}_n$, such that $\mathcal{S}$*simulates* every one of the $\mathcal{S}_n$ and they converge to $\mathcal{S}$ in the metrics of the last chapter. Thus, "in the limit", the $\mathcal{S}_n$ are behaviourally equivalent to the original process. In fact there is an even better property: for every modal formula ϕ satisfied by $\mathcal{S}$ there is an n such that $\mathcal{S}_n$ satisfies ϕ. One can therefore verify whether $\mathcal{S}$ satisfies a formula of interest, by checking whether that formula is satisfied by an appropriately chosen $\mathcal{S}_n$. Best of all, one can even "adapt" the approximation scheme to a set of formulas $\mathcal{F}$ in such a way that we produce a finite-state approximant $\mathcal{Q}$ with the property that for any formula $\phi \in \mathcal{F}$, $\mathcal{S}$ satisfies ϕ if and only if $\mathcal{Q}$ does.

9.1 An Explicit Approximation Construction

In this section we present the first approximation construction; this appeared in [DGJP00] with the full version of the paper appearing in [DGJP03]. It is based on partitioning the state space of the **LMP** into finitely many blocks. One gets better and better approximations by refining these blocks. Every approximant is simulated by the original system, so every formula satisfied by an approximant is also satisfied by the **LMP** being approximated. However, more is true, as we have noted above. Every formula that is satisfied by the "big" system is also satisfied by *some* approximant.

For any labelled Markov process $\mathcal{S}$, we build a sequence of *finite acyclic labelled Markov processes* (FAMPs for short; finite-state processes in which the transition graph is acyclic) $\{\mathcal{S}_i\}$ such that:

(1) For any formula ϕ satisfied by $\mathcal{S}$, there is an i such that $\mathcal{S}_i$ satisfies ϕ,
(2) $\mathcal{S}_i$ is simulated by $\mathcal{S}$, for all i.

In fact, the approximants form a chain in the simulation ordering and the least upper bound of this chain is $\mathcal{S}$. Furthermore, the sequence $\mathcal{S}_i$ forms a Cauchy sequence in the metric of the previous chapter. The limit of this sequence is also $\mathcal{S}$. Thus, there are two senses in which the approximations converge to $\mathcal{S}$. We will tell the simulation story in the next chapter when we give a domain-theoretic account of approximation.

9.1.1 *Finite-state approximation: first attempt*

There are two parameters to the approximation: one is a natural number n, and the other is a positive rational ϵ. The number n gives the number of successive transitions possible from the start state. The number ϵ measures the accuracy with which the probabilities approximate the transition probabilities of the original process. In order to obtain a countable family, we will require that ϵ be a rational number, but, of course, in reality any real number can be chosen for ϵ.

Given a labelled Markov process $\mathcal{S} = (S, i, \Sigma, \tau)$, an integer n and a rational number $\epsilon > 0$, we define $\mathcal{S}(n, \epsilon)$ to be an n-step unfolding approximation of $\mathcal{S}$. Its state-space is divided into $n+1$ levels which are numbered $0, 1, \ldots, n$. At each level, say n, the states of the approximant is a partition of S; these partitions correspond to the equivalence classes corresponding to an approximation to bisimulation. The initial state of $\mathcal{S}(n, \epsilon)$ is at level

n; transitions only occur between a state of a given level to a state of one lower level. Thus, in particular, states of level 0 have no outgoing transitions. This makes the approximants acyclic. This sometimes leads to silly situations where a loop is unwound and the finite-state approximation is too large. We will modify the construction to take care of this later.

In the following we omit the curly brackets around singleton sets.

Definition 9.1 Let (S, i, Σ, τ) be a labelled Markov process, $n \in \mathbf{N}$ and ϵ a positive rational. We denote the finite-state approximation by $\mathcal{S}(n, \epsilon) = (P, p_0, \rho)$ where P is a subset of $\Sigma \times \{0, \dots, n\}$. It is defined as follows, for $n \in \mathbf{N}$ and $\epsilon > 0$. $\mathcal{S}(n, \epsilon)$ has $n+1$ levels. States are defined by induction on their level. Level 0 has one state $(S, 0)$. Now, given the sets from level l, we define states of level $l+1$ as follows. Suppose that there are m states at level l, we partition the interval $[0, 1]$ into intervals of size ϵ/m. Let $(B_j)_{j \in I}$ stand for this partition; i.e. for $\{\{0\}, (0, \epsilon/m], (\epsilon/m, 2\epsilon/m], \dots\}$. States of level $l+1$ are obtained by the partition of S that is generated by the sets $\tau_a(\cdot, C)^{-1}(B_j)$, for every set C corresponding to state at level l and every label $a \in \{a_1, \dots, a_n\}$, $i \in I$. Thus, if a set X is in this partition of S, $(X, l+1)$ is a state of level $l+1$. Transitions can happen from a state of level $l+1$ to a state of level l, and the transition probability function is given by

$$\rho_a((X, k), (B, l)) = \begin{cases} \inf_{t \in X} \tau_a(t, B)) & \text{if } k = l+1, \\ 0 & \text{otherwise.} \end{cases}$$

The initial state p_0 of $\mathcal{S}(n, \epsilon)$ is the unique state (X, n) such that X contains i, the initial state of $\mathcal{S}$.

If $B = \cup B_j$, is a (finite and disjoint) union of sets at level l, we will write (B, l) for the set $\{(B_1, l), (B_2, l), \dots\}$ of all of the corresponding states, and by extension, we will write $\rho_a((X, l+1), (B, l))$ to mean $\sum_{j \in I} \rho_a((X, l+1), (B_j, l))$. If $s \in S$, we denote by (X_s, l) the unique state at level l such that $s \in X_s$.

The following lemma is a trivial but useful result. It is a consequence of the construction of finite approximants and uses crucially the fact that the partition of $[0, 1]$ takes into account the number of states m at the preceding level.

Lemma 9.1 *Let $\mathcal{S}$ be a labelled Markov process, and $s \in S$. In $\mathcal{S}(n, \epsilon)$, if B is a finite union of sets appearing at level l, then*

$$0 < \tau_a(s, B) - \rho_a((X_s, l+1), (B, l)) \leq \epsilon.$$

Proof. Let $(X, l+1)$, (B_j, l), $j = 1, \ldots, k$ be states of $\mathcal{S}(n, \epsilon)$. Let m be the number of states at level l. Then for all $s, t \in X$ we have

$$|\tau_a(s, B_j) - \tau_a(t, B_j)| < \epsilon/m,$$

because of the way S is partitioned on level $l+1$. Thus we have

$$\begin{aligned} |\tau_a(s, B) - \rho_a((X, l+1), (B, l))| &= |\tau_a(s, B) - \sum_{j=1}^{k} \inf_{t \in X} \tau_a(t, B_j))| \\ &\leq \sum_{j=1}^{k} |\tau_a(s, B_j) - \inf_{t \in X} \tau_a(t, B_j))| \\ &\leq \sum_{j=1}^{k} \epsilon/m \\ &\leq \epsilon. \end{aligned}$$

□

The next theorem says that every state (X, l) in $\mathcal{S}(n, \epsilon)$ is simulated in $\mathcal{S}$ by every state $s \in X$.

Theorem 9.1 *Every labelled Markov process $\mathcal{S}$ simulates all its approximations of the form $\mathcal{S}(n, \epsilon)$. More precisely, every state (X, l) of $\mathcal{S}(n, \epsilon)$ $(l \leq n)$ is simulated in $\mathcal{S}$ by every $s \in X$.*

Proof. Let $\mathcal{S}(n, \epsilon) = (P, p_0, \rho)$ and $\mathcal{U} = (U, u_0, \Omega, \nu)$ be the direct sum of $\mathcal{S}(n, \epsilon)$ and $\mathcal{S}$. Now let R be the reflexive relation on U relating a state (X, l) from $\mathcal{S}(n, \epsilon)$ to every state $s \in X$ from $\mathcal{S}$. We prove that R is a simulation. Consider two related states, (X, l) and $s \in X$ and let $Y \in \Omega$ be R-closed, that is, $Y \cap S \in \Sigma$ and $R(Y \cap P) \subseteq Y$. The only positive transitions from (X, l) are to states of the form $(B, l-1)$ so we can assume that $Y \cap P$ is a union B of states of level $l-1$. Now observe that $R((B, l-1)) \cap S = B$ and by the preceding lemma we have:

$$\begin{aligned} \nu_a((X, l), (B, l-1) \cup B) &= \rho_a((X, l), (B, l-1) \\ &\leq \tau_a(s, B) \\ &= \nu_a(s, (B, l-1) \cup B), \end{aligned}$$

and hence the result follows. □

The next theorem links the approximation with the logic. The proof is omitted; in the next section we prove a theorem that subsumes this one. It shows that the approximation process eventually captures every formula satisfied by the system being approximated. This also shows that the formulas capture "finite pieces of information" about the system.

Theorem 9.2 *If a state $s \in S$ satisfies a formula $\phi \in \mathcal{L}_\vee$, then there is some approximation $\mathcal{S}(n, \epsilon)$ such that $(X_s, n) \models \phi$.*

We now show that one can reconstruct the original process – more precisely, a bisimulation equivalent of the original process – from the approximants $\mathcal{S}(n, \epsilon)$. We do not reconstruct the original state space but instead the bisimulation equivalence classes of it. That is why we state in the theorem that one has to start with a bisimulation-collapsed version of the system.

Theorem 9.3 *Let (S, i, Σ, τ) be a labelled Markov process that is maximally collapsed, that is, $S = S/_{\approx}$. If we are given all finite-state approximations $\mathcal{S}(n, \epsilon)$, we can recover (S, i, Σ, τ).*

Proof. We can recover the state space as a set trivially by taking the union of states at any level of any approximation. We know, from the fact that $\mathcal{S}$ is maximally collapsed, that Σ is generated by the sets of the form $[\![\phi]\!]$. This is a consequence of the quotient Lemma 7.16 appearing in the proof of the logical characterisation of bisimulation. Moreover, in the proof of Theorem 9.2 above, we proved that for every $\mathcal{S}$ and every formula ϕ, there exist states, in some approximation $\mathcal{S}(n, 1/2^n)$, such that the union of the sets X_n representing these states is exactly the set of states in $\mathcal{S}$ that satisfy ϕ; i.e. $\cup_{n \geq l} X_n = [\![\phi]\!]_{\mathcal{S}}$. These two facts imply that

$$\mathcal{B} := \{B : (B, l) \in \mathcal{S}(l, 1/2^n) \text{ for } l, n \in \mathbf{N}\}$$

generates Σ (obviously, $\mathcal{B} \subseteq \Sigma$).

The main difficulty is that we have to recover the transition probability function. To do so, let $\mathcal{F}(\mathcal{B})$ be the set containing finite unions of sets in $\mathcal{B}$. We first argue that $\mathcal{F}(\mathcal{B})$ forms a field, then we define $\nu_a(s, \cdot)$ on it and we show that $\nu_a(s, \cdot)$ and $\tau_a(s, \cdot)$ agree on it for all $s \in S$. This will imply that $\nu_a(s, \cdot)$ is finitely additive on $\mathcal{F}(\mathcal{B})$ and hence can be extended uniquely to a measure on Σ, and hence ν_a and τ_a agree on $S \times \Sigma$, as desired.

We now show that $\mathcal{F}(\mathcal{B})$ forms a field. It is obviously closed under finite unions. To see that it is also closed under intersection and complementation, note that if $(C, n) \in \mathcal{S}(n, \epsilon)$, then for all $m > n$ and all δ such that ϵ is an integer multiple of δ, C is a union of a family of sets C_i such that $(C_i, m) \in \mathcal{S}(m, \delta)$. This is clear from the fact that the construction proceeds by splitting existing blocks. From this observation it is clear that we have closure under intersections because given two sets we can move to a stage of the approximation process that refines both of them. At this stage the intersection of the two sets that we started out with is clearly a union of the sets at the present stage.

Now let $C \in \mathcal{F}(\mathcal{B})$, $s \in S$, $a \in \mathcal{A}$ and let

$$\nu_a(s, C) := \sup_{n,\epsilon} \sum_{\substack{B \subseteq C \\ (B,n-1) \in \mathcal{S}(n,\epsilon)}} \rho_a((X_s, n), (B, n-1)).$$

We prove that $\nu_a(s, \cdot)$ and $\tau_a(s, \cdot)$ agree on $\mathcal{F}(\mathcal{B})$ for all $s \in S$. Obviously, $\nu_a(s, C) \leq \tau_a(s, C)$ for $C \in \mathcal{F}(\mathcal{B})$. The reverse inequality follows from Lemma 9.1:

$$\begin{aligned}
\sup_{n,\epsilon} \sum_{\substack{B \subseteq C \\ (B,n-1) \in \mathcal{S}(n,\epsilon)}} \rho_a((X_s, n), (B, n-1)) &= \sup_{n,\epsilon} \rho_a((X_s, n), (\cup B, n-1)) \\
&\geq \sup_{n,\epsilon} (\tau_a(s, \cup B) - \epsilon) \\
&\geq \sup_{(n,\epsilon) \in I} (\tau_a(s, C) - \epsilon) \\
&= \tau_a(s, C),
\end{aligned}$$

where I is the set of pairs (n, ϵ) such that in $\mathcal{S}(n, \epsilon)$, level n contains a partition of C (note that there are arbitrary small ϵ's that are involved in I). This concludes the proof that ν and τ agree and we are done. □

We conclude the discussion of the basic approximation scheme by showing that – in the metrics introduced in the last chapter, with $c < 1$ – the approximants converge to the labelled Markov process being approximated.

In order to prove convergence of the approximants we start with the following lemma. Many of the lemmas are routine and their proofs are omitted, a complete version of all the proofs can be found in Desharnais's thesis [Des99].

Lemma 9.2 *If $\mathcal{S}$ involves a finite number of labels, $\mathcal{S}(n, c^n/n)$ converges to $\mathcal{S}$ in the metric d_c with $c < 1$.*

The condition $c < 1$ is important in the calculation.

Lemma 9.3 *Given any process of the form $\mathcal{S}(n, \epsilon)$ we can construct a sequence of rational trees $\mathcal{T}_i$ such that $\mathcal{T}_i$ is strictly simulated by $\mathcal{T}_{i+1}$ and all of them are strictly simulated by $\mathcal{S}(n, \epsilon)$ and with* $\lim_{i \to \infty} d(\mathcal{T}_i, \mathcal{S}(n, \epsilon)) = 0$.

Proof. Given a finite acyclic process like $\mathcal{S}(n, \epsilon)$ we can construct a finite depth tree, say $\mathcal{T}$, that is bisimilar to it by duplicating states as necessary. The transition probabilities of this tree will not necessarily be rational numbers. We can construct the required family of trees by making all the $\mathcal{T}_i$ have the same shape as $\mathcal{T}$ but by choosing the transition probabilities in the $\mathcal{T}_i$ to be rational numbers converging to the corresponding transition probabilities of $\mathcal{T}$. Since these probabilities are all strictly increasing we get the desired strict simulation. The convergence is immediate from the definition of the metric. □

The theorem below actually says a little more, the rational trees are dense with the metric topology.

Theorem 9.4 *For all $c \in (0, 1]$, the metric d^c yields a separable metric space.*

Proof. We show that the rational trees form a countable dense subset. Given any process $\mathcal{S}$ we have a countable family of finite approximations given by $\mathcal{S}(n, 2^{-n})$. For each of these finite approximations we have a countable sequence of rational trees, $\mathcal{T}_j^{(n)}$ that converge to it by the previous lemma. This doubly indexed family of rational trees forms a directed set so we can extract a countable sequence of rational trees that converge to $\mathcal{S}$□

Thus we have a situation analogous to the rationals where there is a countable family that serves to approximate all the processes as limits of Cauchy sequences. What we do not know is whether the metric space is complete; in other words we do not know whether we have a Polish space. The following corollary makes precise the claim made above about convergence.

Corollary 9.1 *If $\mathcal{S}$ involves a finite number of labels, $\tilde{\mathcal{S}}_{n,c^n/n}$ converges to $\mathcal{S}$ in the metric d^c with $c < 1$.*

9.2 Dealing With Loops

The above unfolding construction is very simple but sometimes it does very silly things. Consider a one-state system with a probability 1 transition from the state back to itself with a single label a. The above construction will take this nice finite-state system and "unwind" it to produce a sequence of increasingly long chains. These will indeed be finite-state approximations of the original system but the original system was already a very simple finite-state system. We need to fix the above construction to deal with loops.

The state-space is constructed in the same manner but there will be "more" transitions possible, that is, transitions that produce cycles. There are two parameters to the approximation: one is a natural number n, and the other is a positive rational ϵ. The number ϵ, as before, measures the accuracy of the approximation. Since the state-space is constructed in the same way as above, all the results that we had about the state-space of the previous construction can be used here. The difference between the two constructions lies in the transitions.

Given a labelled Markov process $\mathcal{S} = (S, \Sigma, h)$, a natural number n and a rational number $\epsilon > 0$, we define $\mathcal{S}^*(n, \epsilon)$ as an n-step unfolding approximation of $\mathcal{S}$. Its state space is divided into $n+1$ levels which are numbered $0, 1, \ldots, n$. At each level, say n, the states of the approximant are the blocks in a partition of S. The initial state of $\mathcal{S}^*(n, \epsilon)$ is at level n and transitions only occur between a state of one level to a state of one lower level *or between states of the same level.* Thus, in particular, the unique state of level 0 either has no outgoing transitions or has a transition to itself. The main difference with what we did before is that we now permit transitions to states at the same level. Thus, in particular, the transition graph could be cyclic and no longer has a well defined depth. The parameter "n" refers to the extent to which we probe the process by our observations.

Definition 9.2 Let (S, i, Σ, h) be a labelled Markov process, $n \in \mathbf{N}$ and ϵ a positive rational. We denote the finite-state approximation by $\mathcal{S}^*(n, \epsilon) = (P, p_0, \wp P, \rho)$ where

- P is a subset of $\Sigma \times \{0, \ldots, n\}$; the numbers from 0 to n correspond to the level of the states. States are defined by induction on their level.
  – At level 0 there is one state $(S, 0)$.
  – Now, given the states $(C_1, l), (C_2, l), \ldots, (C_m, l)$ at level l, we define states of level $l+1$ as follows. Let $(B_j)_{j \in I}$ be the partition

  $$\{\{0\}, (0, \epsilon/m], (\epsilon/m, 2\epsilon/m], \ldots\}$$

  of the interval $[0, 1]$ into intervals of size ϵ/m. States of level $l+1$ are obtained by forming the coarsest common refinement of the partition $\{C_i\}_{i=1}^m$ and the partition generated by the sets $h_a(\cdot, C_i)^{-1}(B_j)$, for every set C_i and every label $a \in \{a_1, \ldots, a_n\}$, $j \in I$. If a set X is in this partition of S, $(X, l+1)$ is a state of level $l+1$.

- The initial state p_0 of $\mathcal{S}^*(n,\epsilon)$ is the unique state (X,n) such that X contains i, the initial state of $\mathcal{S}$.
- Transitions can happen between states of the same level, or from a state to a state of the preceding level, and the transition probability function is given as follows. Let $(X,l+1),(Y,l+1),(Z,l)$ be states of level $l+1$ and l, where $l \geq 0$. Then

$$\rho_a((X,l+1),(Y,l+1)) = \inf_{x\in X} h_a(x,Y)$$

$$\rho_a((X,l+1),(Z,l)) = \inf_{x\in X} h_a(x,Z) - \sum_{i=1}^{k} \rho_a((X,l+1),(Z_i,l+1))$$

  where $\{Z_i\}_{i=1}^k$ is the unique partition of Z such that $(Z_i,l+1)$ is a state for every i. Unspecified transitions are given the value 0.

The partition of S at level $l+1$ is defined in such a way that every state $x \in X$ (where X is a member of the partition) has probability within ϵ/m to every set in the partition of level l (not necessarily true for transitions to states of level $l+1$). Intuitively, transitions are filled as follows: from a given state $(X,l+1)$, transitions to states at the same level are given the maximum probability possible (staying below all *simulating* states $x \in X$). This would not be sufficient to guarantee that the transition be close to the corresponding transition of $\mathcal{S}$ because the partition of level $l+1$ is constructed with respect to states of level l. Since this condition is essential to preserve the accuracy of the approximation – and the statement of the lemma below reflects this – we complete the probability by adding transitions to states (Z,l).

Notation

If $s \in S$, we denote by (X_s,l) the unique state at level l such that $s \in X_s$. We will write (Y,l) for the set $\{(Y_1,l),(Y_2,l),\ldots\}$, where $Y = \cup Y_j$; in this case, we often say that Y is a *union of sets at level l* and that the Y_i's *correspond to states of level l*. By extension, we will write $\rho_a((X,l+1),(Y,l))$ to mean $\sum_j \rho_a((X,l+1),(Y_j,l))$. The same notation will be used when we work with states of consecutive levels corresponding to the same subset of S: for example, we will write $(Y, l \cup l+1)$ to mean $\{(Y_1,l),(Y_2,l),\cdots,(Y_1',l+1),(Y_2',l+1),\cdots\}$, with $\cup Y_i = \cup Y_i' = Y$. Note that every set of level $l-1$ is a union of sets of level l because the partition of S at level l is a refinement of the partition at level $l-1$.

The following lemma is the analogue of Lemma 9.1.

Lemma 9.4 *Let $\mathcal{S}$ be a labelled Markov process, and $s \in S$. In $\mathcal{S}^*(n,\epsilon)$, if Y is a union of sets appearing at level l, then $0 < h_a(s,Y) - \rho_a((X_s, l+1),(Y, l \cup l+1)) \leq \epsilon$.*

Proof. The first inequality is trivial. Before proving the second one, note that the lemma is not necessarily true if Y is a union of sets appearing at the same level as $(X_s, l+1)$.

Let $s \in S$ and $(X_s, l+1)$, $(Y_i, l+1), (Y'_j, l)$, $i=1,\dots,k$, $j=1,\dots,k'$ be states of $\mathcal{S}^*(n,\epsilon)$ such that $Y = \cup_{i=1}^{k} Y_i = \cup_{j=1}^{k'} Y'_j$. Let m be the number of states at level l. Then for all $j = 1,\dots,k'$ and $t \in X_s$ we have

$$|h_a(s, Y'_j) - h_a(t, Y'_j)| < \epsilon/m,$$

because of the way S is partitioned on level $l+1$. Moreover, we have

$$\begin{aligned}
&\rho_a((X_s, l+1),(Y, l \cup l+1)) \\
&\qquad = \rho_a((X_s, l+1),(Y, l)) + \rho_a((X, l+1),(Y, l+1)) \\
&\qquad = \sum_{j=1}^{k'} \rho_a((X_s, l+1),(Y'_j, l)) + \sum_{i=1}^{k} \rho_a((X_s, l+1),(Y_i, l+1)) \\
&\qquad = \sum_{j=1}^{k'} \inf_{s \in X_s} h_a(s, Y'_j))
\end{aligned}$$

and hence

$$\begin{aligned}
&|h_a(s,Y) - \rho_a((X_s, l+1),(Y, l \cup l+1))| \\
&\qquad = |\sum_{j=1}^{k'} h_a(s, Y'_j) - \sum_{j=1}^{k'} \inf_{s \in X_s} h_a(s, Y'_j)| \\
&\qquad \leq \sum_{j=1}^{k'} |h_a(s, Y'_j) - \inf_{s \in X_s} h_a(s, Y'_j)| \\
&\qquad \leq \sum_{j=1}^{k'} \epsilon/m \\
&\qquad \leq \epsilon.
\end{aligned}$$

□

Since every transition probability of $\mathcal{S}^*(n,\epsilon)$ is smaller than in the corresponding transition in $\mathcal{S}$, then every state (X, l) in $\mathcal{S}^*(n,\epsilon)$ is simulated by every state $s \in X$ in $\mathcal{S}$.

Theorem 9.5 *Every labelled Markov process $\mathcal{S}$ simulates all its approximations of the form $\mathcal{S}^*(n,\epsilon)$. More precisely, every state (X,l) of $\mathcal{S}^*(n,\epsilon)$ $(l \le n)$ is simulated in $\mathcal{S}$ by every $s \in X$.*

This is the analogue of Theorem 9.1. The proof is omitted; it is conceptually the same as above but the bookkeeping makes it harder to read.

The next theorem is the analogue of Theorem 9.2. The proof is exactly the same as for the previous version of the construction except for the very last sequence of inequalities, which is adapted to the fact that transitions can happen between states of the same level. Notice that here we use a semantics for $\mathcal{L}_\vee$ with strict inequality in the modal formula.

Theorem 9.6 *If a state $s \in S$ satisfies a formula $\phi \in \mathcal{L}_\vee$, then there is some approximation $\mathcal{S}^*(n,\epsilon)$ such that $(X_s, n) \models \phi$.*

Proof. The proof is by induction on the structure of formulas. We prove the following stronger induction hypothesis. We prove that for all formulas ϕ there is an increasing sequence $(X_n)_{n \ge depth(\phi)}$ of sets in Σ which satisfy:

(i) $\cup_{n \ge depth(\phi)} X_n = [\![\phi]\!]_{\mathcal{S}}$;
(ii) $X_n = \cup_{s \in X_n} C_s$, where $(C_s, l) \in \mathcal{S}^*(n, 1/2^n)$ and $l \ge depth(\phi)$;
(iii) the states (C_s, l) satisfy ϕ in $\mathcal{S}^*(n, 1/2^n)$.

It is obvious for T with $X_n = S$ for all n.

Consider $\phi = \phi_1 \wedge \phi_2$. Assume the claim is true for ϕ_j, $j = 1, 2$. Let $(X_n^j)_{n \ge depth(\phi_j)}$ be the sequence for ϕ_j. Now define for $n \ge depth(\phi)$, the sequence

$$X_n = X_n^1 \cap X_n^2.$$

Note that this is an increasing sequence of sets in Σ. We first prove (i): for all $s \models \phi$, there is some n such that $s \in X_n$. Choose $n = \max(n_1, n_2)$ where n_j is such that $s \in X_{n_j}^j$. Now for (ii) and (iii), let $s \in X_n$, for a fixed $n \ge depth(\phi)$. Then because all states (C_s, l) satisfy ϕ_j and $C_s \subseteq X_n^j$, we have $(C_s, l) \models \phi_1 \wedge \phi_2$ and $X_n = \cup_{s \in X_n} C_s$. The proof for the case $\phi_1 \vee \phi_2$ is similar.

Consider $\phi' = \langle a \rangle_q \phi$, and assume the claim is true for ϕ. Let $d = depth(\langle a \rangle_q \phi)$, $\epsilon_n = 1/2^n$ and let $(X_n)_{n \ge d-1}$ be the sequence for ϕ.

Now define for $n \ge d$, the sequence

$$Y_n = \cup\{C : (C,d) \in \mathcal{S}^*(n,\epsilon_n), \text{ and } \forall s \in C, h_a(s, X_n) > q + \epsilon_n\}.$$

This is an increasing sequence of sets in Σ because if $(C,d) \in \mathcal{S}^*(n,\epsilon_n)$ and $C \subseteq Y_n$, then for all $s \in C$ we have $h_a(s, X_{n+1}) \ge h_a(s, X_n) \ge$

$q + \epsilon_n$. Moreover, if (C', d) is a state of $\mathcal{S}^*(n, \epsilon_{n+1})$ and $s, t \in C'$, then $h_a(t, X_{n+1}) > h_a(s, X_{n+1}) - \epsilon_{n+1} \geq q + \epsilon_n - \epsilon_{n+1} = q + \epsilon_{n+1}$.

We now prove (i), that is, for all $s \models \phi'$, there is some n such that $s \in Y_n$. So assume $h_a(s, [\![\phi]\!]) > q$. Then there is some n such that $h_a(s, X_n) - q > 2\epsilon_n$ because $h_a(s, \cdot)$ is a measure and X_n is an increasing sequence which converges to $[\![\phi]\!]$ and ϵ_n $(= 1/2^n)$ is decreasing to 0. Now since X_n is a union of states of level $l - 1 \geq d - 1$, then for every $t \in C_s$, with (C_s, l) a state of $\mathcal{S}^*(n, \epsilon_n)$ we have

$$|h_a(s, X_n) - h_a(t, X_n)| < \epsilon_n$$

and hence $h_a(t, X_n) - q > \epsilon_n$. Thus $C_s \subseteq Y_n$ and (i) and (ii) are proved. Note that the inequality sign in the meaning of the modal formula was crucial to this part of the proof.

We now prove (iii). Let $s \in Y_n$, for a fixed $n \geq d$. Then because all states $(X, l - 1)$, where $X \subseteq X_n$ and $l - 1 \geq d - 1$, satisfy ϕ and by Lemma 9.1, we have

$$\begin{aligned} \rho_a((C_s, l), ([\![\phi]\!]_{\mathcal{S}^*(n,\epsilon_n)}, l \cup l - 1)) &\geq \rho_a((C_s, l), (X_n, l \cup l - 1)) \\ &\geq h_a(s, X_n) - \epsilon_n \\ &> q + \epsilon_n - \epsilon_n = q. \end{aligned}$$

From this it follows that $(C_s, l) \models \phi'$ for all $l \geq d$; this is what we needed to establish (iii). □

The following results shows that a finite process is eventually approximated by itself. This is the main reason why we have introduced this new construction.

Corollary 9.2 *For every finite process there exists a bisimilar approximation.*

Proof. Since the process $\mathcal{S}$ is finite, the partition at the highest level of $\mathcal{S}^*(n, 1/2^n)$ must stabilise when n increases. In fact, it must converge to the bisimulation equivalence classes. Indeed, if two states are not bisimilar they must be distinguished by a formula ϕ. Then by the (proof of the) previous theorem there is some n such that the two states are not in the same set of $\mathcal{S}^*(n, 1/2^n)$. Thus the partition at the highest level corresponds exactly to the bisimulation equivalence classes. By construction of approximants, transitions from states of this level will only happen to states of this same level and hence the result. □

Corollary 9.3 *Let $\mathcal{S}$ be an LMP. Then for $c < 1$ we have*

$$d^c(\mathcal{S}, \mathcal{S}^*(n, 1/2^n)) \longrightarrow 0$$

and it is also true for $c = 1$ if the set of infinite sequences of non-bisimilar states starting in the initial state of $\mathcal{S}$ is of measure 0.

9.3 Adapting the Approximation to Formulas

In the last section we showed that any logical formula of interest could be checked on a suitable finite approximation. How far does one have to approximate in order to be sure that the formula of interest has been captured? In other words, suppose that we are interested in knowing if a certain **LMP** $\mathcal{S}$ satisfies a formula ϕ. If at a certain stage of the approximation we find that ϕ does not hold, what can we conclude? Does it mean that $\mathcal{S}$ does not satisfy ϕ? How do we know whether or not a later approximation will show that $\mathcal{S}$ does satisfy ϕ?

If the logic is simple enough – like the logic $\mathcal{L}_\vee$ – then one can estimate how far one has to push n to ensure that a formula like ϕ holds in $\mathcal{S}$. This is because the nesting depth of the modal operator counts how many steps of $\mathcal{S}$ are being probed and thus one can relate this to the approximation parameter n.

What if the logic is more complicated? Danos and Desharnais [DD03b; DD03a; DDP04] show how to "orient" the approximation process so that given a suitable family of formulas $\mathcal{F}$, in a logic with fixed-point operators[1] – hence with all the usual temporal connectives – one can construct an approximation such that a formula of $\mathcal{F}$ is satisfied by $\mathcal{S}$ if and only if it is satisfied by the approximant.

This is very nice indeed, but there is one small problem. The resulting approximations are not probabilistic automata at all. The transition "probabilities" are not additive, i.e. they are not probabilities. They are not, however, completely arbitrary. They are *super additive*; this means that the "probability" of the union of two disjoint sets is less than or equal to the sums of the individual probabilities. These are not measures at all, but are closely related to mathematical objects called capacities [Cho53] and obey quite a number of nice mathematical properties.

[1] The formulas in this logic are presented as automata; it is well-known that modal logics and automata are practically the same thing.

Why does this happen? It is essentially because the approximation process that we use takes infima of the real probabilities to estimate the approximate probabilities. This guarantees that the approximation is simulated by the original system. However, it may lead to a serious underestimation of the transition probabilities.

Chapter 10

Approximating the Approximation

The approximation schemes described in the last chapter have nice mathematical properties, but can we really use them in practice? A number of problems arise when one attempts the approximation construction. The most vexing is that the number of states rises very rapidly. The second problem is that many of the states, which are subsets of the original continuous state space, become very small. The third problem is that these subsets may be very "nasty"; they are only required to be measurable. In this section we describe work reported in [BCFPP05] where Monte Carlo techniques were used to alleviate some of these problems. There are still, however, many obstacles to using these techniques in practice and it is the subject of active investigation.

The first question one must face before doing computation in a continuous state space is the representation problem: how does one express transition probability kernels assuming an uncountable range of values? In many cases we have a "canonical" probability measure μ on (S, Σ). This is, for example, Lebesgue measure on a subset of $\mathbf{R}^n$ or, if the state space is a manifold, the induced geometric measure. In this case one can use:

Definition 10.1 *A family of sub-probability density functions* $f_a : S^2 \longrightarrow [0, \infty)$, $s \in S, a \in A$, which is simply a family of $(\mathfrak{M} \otimes \mathfrak{M})$-measurable functions such that

$$\int_S f_a(s_0, \cdot) d\mu \leq 1 \; \forall s_0 \in S, a \in A.$$

Then, the kernels are given by:

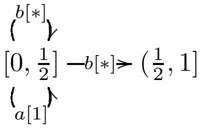
$$\tau_a(s_0, M) := \int_M f_a(s_0, \cdot) d\mu \; \forall M \in \mathfrak{M}, a \in A, s_0 \in S.$$

It is not hard to show that τ_a is then a labelled probability kernel (the fact that $\tau_a(s,\cdot)$ is a measure follows using the monotone convergence theorem, the measurability of $\tau_a(\cdot, M)$, using Fubini's theorem). We will denote this construction by $d\tau_a(s,\cdot) = f_a(s,\cdot)d\mu$. Recall that this representation is possible iff $\tau_a \ll \mu$ (by the Lebesgue-Radon-Nikodym theorem).

Example 10.1 We now show a toy example of this construction that will be used later to test our algorithms. Consider a pair of 2-dimensional aquaria, arranged side by side horizontally. The first aquarium has horizontal coordinates $[0, \frac{1}{2}]$ and the second, $(\frac{1}{2}, 1]$. We are interested in the evolution of the horizontal position of a stochastic fish that starts its life in the first aquarium and has a choice of two actions:

- *swim* will change the position of the fish in its aquarium. The new position is drawn uniformly from the interval $[0, \frac{1}{2}]$.
- *jump* corresponds to an attempt to jump into the second aquarium. If the fish is at distance d from Aquarium 2, it will fail and fall in the original aquarium with probability $2d$ (the next position will then be drawn uniformly from the interval $[\frac{1}{2} - d, \frac{1}{2}]$). If the fish succeeds, its new position is drawn uniformly from the interval $(\frac{1}{2}, 1]$. Unfortunately, the fish does not know that the second aquarium is filled with a liquid fatal for its metabolism. The death of the fish is modelled by disabling both actions in the second aquarium.

Schematically:



where the label $a[p]$ on an edge (s_i, s_j) denotes that the probability to transition from s_i to s_j is p, given that action a is selected.

Note that the probability distribution induced by selecting action b is different for each state in $[0, \frac{1}{2}]$ from which the action is taken. This is denoted by $b[*]$. Hence, this LMP cannot be lumped into a finite state system.

Let $\mu =$ the Lebesgue measure on $[0, 1]$. We obtain easily that the kernels corresponding to the actions described above can be expressed using the following probability density functions:

$$f_{\text{swim}}(x,y) := \begin{cases} 1 \text{ if } x \in [0,\frac{1}{2}] \text{ and } y \in [0,\frac{1}{2}] \\ 0 \text{ otherwise} \end{cases}$$

$$f_{\text{jump}}(x,y) := \begin{cases} 2 \text{ if } x \in [0,\frac{1}{2}] \text{ and } y \in [x,\frac{1}{2}] \\ 4x \text{ if } x \in [0,\frac{1}{2}] \text{ and } y \in (\frac{1}{2},1] \\ 0 \text{ otherwise} \end{cases}$$

Example 10.2 Let us consider now a more realistic situation. Consider the on board flight control system of a Cosmos-3MU launcher, a 2-stage, UDMH-fuelled dispensable rocket often used to send small payloads into Earth orbit [Web].

A hypothetical problem for which the approximation scheme would be useful is the verification and/or evaluation of the effectiveness of flight guidance software for the Cosmos-3MU (In November 2000, and twice in January 2005, the second stage of the launcher failed to form the final orbit because of undiagnosed problems in this system. At least two commissions tried unsuccessfully to isolate the source of this "bug"). Suppose that the main controller must keep the launcher within distance $r_{\max}$ of the ideal trajectory by applying lateral speed corrections. The controller is composed of a sampling loop, the cyclic executive, that applies a thrust towards the ideal trajectory if needed. This loop structure motivates the discrete-time model of the problem. The state space is the cartesian product of the velocity space with the distance-to-trajectory space, $\mathbb{R}^2$ (with $r = -r, v = -v$). The actions are:

- *actuate*, which applies a velocity correction towards the ideal trajectory. Due to the limited precision of these corrections, the result of this action from state (r_0, v_0) is modelled by a bivariate normal distribution centered at $\big(x_0 + \delta(v_0 - a_{\text{impulse}}), v_0 - a_{\text{impulse}}\big)$ with a strong, positive correlation such that the major axis of the elliptic isopleths of the density (that is, the locus of the points in the plane where $f_{\text{actuate}}(x,v) = c$) has a slope of $1/\delta$. The variance parameters can be set using the large amount of flight data available for this type of rocket (more than 400).
- *stay*, which corresponds to the absence of velocity correction. It is modelled by a normal distribution f_{stay}, similar to f_{actuate}, except that the centre is at $\big(x_0 + \delta \cdot v_0, v_0\big)$.

If, at any point in the second stage of the propulsion sequence, the controller fails to maintain the trajectory within distance $r_{\max}$ of the trajectory, a backup system takes control of the guidance. This is represented by disabling all actions. If, on the other hand, the controller successfully maintains the trajectory for 27 minutes, the orbit is reached and a special action, *success*, is enabled which results in a transition to a special *success state*, s_{success}.

Given that the set M, over which the density functions are to be integrated, can be an arbitrary measurable set, the next difficulty is to compute these integrals algorithmically. Numerical integration cannot be applied here because M could be "too nasty" geometrically to allow a nice partitioning. The solution comes from probability theory:

Lemma 10.1 *Let $(\Omega, \mathcal{F}, P)$, $(S, \mathfrak{M}, \lambda)$ be probability spaces. Assume that we can sample the random variables $X_1, X_2, \ldots, X_i : \Omega \longrightarrow S$ identically and independently according to the distribution λ. Then, if $f : S \longrightarrow \mathbb{R}$ is integrable and $M \in \mathfrak{M}$ we have:*

$$\frac{1}{n}\sum_{i=1}^{n}(\chi_M \cdot f) \circ X_i \longrightarrow \int_M f d\lambda \ (a.s.).$$

This standard result is the basis of *Monte Carlo integration.* Its proof is fairly simple:

Proof. We have the following picture:

$$f \circ X_i : (\Omega, \mathcal{F}, P) \ \longrightarrow \ (S, \mathfrak{M}, \lambda) \ \longrightarrow \ (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$$

Using the fact that $X_1, X_2, \ldots$ are independent and that f is measurable, we obtain easily, using Fubini's theorem, that $f \circ X_1, f \circ X_2, \ldots$ are independent. They are also clearly identically distributed and L^1, so we can apply Khinchine's Strong Law of Large Numbers to obtain:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\chi_M \cdot f) \circ X_i &\longrightarrow \int_\Omega (\chi_M \cdot f) \circ X_1 dP \\ &= \int_S \chi_M \cdot f dP_{X_1} \ (a.s.) \\ &= \int_M f dP_{X_1} \\ &= \int_M f d\lambda.\end{aligned}$$

□

The other operations on measurable sets and functions that we encountered in the basic approximation construction are:

(1) given a measurable set M and a measurable function f, compute the infimum of the value attained by f
(2) given two measurable sets M_1, M_2, determine whether their intersection is non-empty (is-$\emptyset$: Sets $\longrightarrow \{0,1\}$, is-$\emptyset(A) = 1$ iff $A = \emptyset$)
(3) given a measurable function f, compute the inverse image of an interval.

In the paper [BCFPP05] it is explained how point (3) can be avoided; some implementation details are discussed as well. The basic idea is that, since we use Monte Carlo integration for the representation of the kernels, the only operation we need to impose on measurable sets is to test membership of a given point (in particular, there is no need for an operation that would express a measurable set as the union of intervals plus a null set, as it would be the case if we were using numerical integration, for instance). The two other points, however, have to be handled. Note that both arbitrary infima and is-$\emptyset$ queries cannot be computed algorithmically in general, so we look for "measure-theoretic" equivalents that are computable (in a randomised computation model). For the case of infs, we use the following concept:

Definition 10.2 Let $(X, \mathfrak{G}, \mu)$ be a measure space. We define the *essential infimum* over $M \in \mathfrak{G}$ of a bounded measurable function $f : (X, \mathfrak{G}) \longrightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ to be:

$$\operatorname{ess}\inf_M f := \sup\left\{a \in \mathbb{R} : \mu\Big(\big\{x \in M : f(x) < a\big\}\Big) = 0\right\}.$$

Now suppose that $\mathcal{S} := (S, \mathfrak{M}, \boldsymbol{\tau})$ is a LMP such that $\tau_a \ll \mu$ for all kernels τ_a in $\boldsymbol{\tau}$ (in which case we will write "$\boldsymbol{\tau} \ll \mu$"), where μ is a measure on S from which we can sample points. In this situation, essential infima have the advantage of being computable:

Lemma 10.2 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and assume that we can sample the random variables $X_1, X_2, \dots, X_i : \Omega \longrightarrow M$, identically and independently according to the distribution μ, where $M \in \mathfrak{M}$ and $\mu(M) > 0$. Then if $f : S \longrightarrow \mathbb{R}$ is bounded and measurable we have:*

$$\min\big\{f \circ X_i : 1 \leq i \leq n\big\} \longrightarrow \operatorname{ess}\inf_M f \text{ (in P-probability).}$$

Proof. First note that $\operatorname{ess}\inf_M f < \infty$ if and only if $\mu(M) > 0$. Let $\epsilon > 0$ be given. By definition of supremum, we have that

$$p_0 := \mu\Big(\big\{s \in M : f(s) - \operatorname{ess}\inf_M f \leq \epsilon\big\}\Big)$$

must be positive. Then for any $i \in \mathbb{N}$,

$$\begin{aligned} &P\Big(\big\{\omega \in \Omega : |f \circ X_i(\omega) - \operatorname*{ess\,inf}_M f| > \epsilon\big\}\Big) \\ &= \mu\Big(\big\{s \in M : |f(s) - \operatorname*{ess\,inf}_M f| > \epsilon\big\}\Big) \\ &= \mu\Big(\big\{s \in M : f(s) - \operatorname*{ess\,inf}_M f > \epsilon\big\}\Big) \\ &= 1 - p_0 < 1. \end{aligned}$$

Hence, by the independence of the X_is, the probability of the intersection of these events as $i \in \{1, 2, \ldots, n\}$ can be made arbitrarily small by by picking n large enough to ensure that we have enough samples of the X_is.□

Similarly, we do not attempt to decide whether sets are empty, rather we restrict ourselves to deciding whether they have measure 0. This is done using the obvious Monte Carlo algorithm which returns true iff all the points sampled from the canonical measure μ do not belong to N. This clearly decides with high probability whether N has μ-measure 0 or not. We shall call this algorithm is-null.

The final theorem states that infs and is-$\emptyset$s can be replaced by ess infs and is-nulls in the approximation algorithm without altering the important properties possessed by the resulting approximations. We omit the proofs here.

Theorem 10.3 *Let $\mathcal{S} = (S, \mathfrak{M}, \boldsymbol{\tau})$ be a LMP, $\boldsymbol{\tau} := \{\tau_a : S \times \mathfrak{M} \longrightarrow [0, 1], a \in A\}$, and $(S, \mathfrak{M}, \mu)$ be a probability measure from which we can sample points and such that $\boldsymbol{\tau} \ll \mu$. Assume also that the start state of $\mathcal{S}$ is μ-randomly selected (A property that is modelled in the following way: we add a state s_0 to S, which will be the starting state. The transition from s_0 to S is μ for all $a \in A$, and the transitions from $s' \in S$ to s_0 are all set to 0).*

Then for all $\epsilon \in \mathbb{Q}, \epsilon > 0$, $n \in \mathbb{N}$, the Monte Carlo rational approximation $\tilde{\mathcal{Q}}_{\epsilon,n}$ *– i.e. the basic approximation construction with* inf*s replaced by ess*inf*s and is-$\emptyset$ replaced by is-null – is computable and has the following properties:*

(1) every state (X, l) of $\tilde{\mathcal{Q}}_{\epsilon,n}$ is simulated in $\mathcal{S}$ by every $s \in X$,
(2) if a state $s \in S$ satisfies a formula $\phi \in \mathcal{L}_0$, then there is some approximation $\tilde{\mathcal{Q}}_{\epsilon_0,n_0}$ such that $(X_s, n) \models \phi$,
(3) given $c \in (0, 1)$, let $\tilde{\mathcal{Q}}_n$ be $\tilde{\mathcal{Q}}_{\epsilon,n}$ with $\epsilon = c^n/n$. Then $\tilde{\mathcal{Q}}_n$ converges to $\mathcal{S}$ with respect to the metric d^c.

Chapter 11

A Domain of Labelled Markov Processes

The approximation concepts that we have described in the last few chapters use ideas from probability and are based on combinatorial or measure-theoretic arguments. There is, however, a very pleasing domain-theoretic version of the approximation theory. In fact the logical view of domains advocated by Abramsky [Abr91b] yields not only the logical characterisation of bisimulation but also the logical characterisation of simulation.

Furthermore, in a suitable category of domains one can construct a *universal* LMP by solving an appropriate domain equation. The solution of this recursive domain equation yields a dcpo that closely resembles the domain of synchronisation trees constructed by Abramsky [Abr91a]. The order in the domain corresponds to simulation, the way-below relation corresponds to strict simulation and equality corresponds to bisimulation. Using the "domain logic" it it possible to show that simulation in the domain is characterised by $\mathcal{L}_\vee$.

11.1 Background on Domain Theory

There are many good references for domain theory but most of them appear in textbooks devoted to programming language theory where the emphasis is on algebraic domains. A good source for domain theory is Prof. Plotkin's unpublished lecture notes – the so-called Pisa Notes – available from his website. The domains that arise in the present context are *continuous* domains. We review some of the basic definitions. A good treatment of continuous domains appears in *Continuous Lattices and Domains* by Gierz *et al.* [GKK$^+$03] which is a revised and updated version of the famous *Compendium of Continuous Lattices* [GKK$^+$80]. The probabilistic analogue of the powerdomain construction that we use here is due to Jones and Plotkin [JP89] and a very nice expository account of this material appears in Claire Jones's thesis [Jon90].

11.1.1 *Basic definitions and results*

Domains are meant to model data types containing *partially defined* elements. They are partially ordered sets (posets) in which the order relation represents information content qualitatively. In other words if $x \leq y$ we should think that y has all the information that x has, and possibly more. A least element – also called a *bottom element* – is meant to represent the completely uninformative value.

Definition 11.1 A **directed set** X in a poset S is a subset of S with the property that for any two elements x, y of X there is an element $z \in X$ that is larger than each of x and y.

Directed sets represent a collection of *consistent* pieces of information. It seems reasonable that we should be able to aggregate consistent information.

Definition 11.2 A partial order is a **dcpo**, if it is closed under limits of directed sets.

We will only consider dcpos with a bottom element. These are the basic ingredients of domain theory.

Definition 11.3 $b \ll x$ ("b is **way-below** x") if for all directed sets X such that $x \sqsubseteq \bigsqcup X$, $(\exists x_i \in X)\ b \sqsubseteq x_i$.

The intuition here is that x contains an "essential" piece of information about y.

Definition 11.4 A dcpo D is **continuous**, if for all $d \in D$, the set $\{b \mid b \ll d\}$ is directed and has lub d. It is ω**-continuous**, if there is a countable subset B such that for all $d \in D$, the set $\{b \in B \mid b \ll d\}$ is directed and has lub d. Such a set B is called a **basis**.

Definition 11.5 $b\Uparrow \overset{d}{=} \{x \mid b \ll x\}$. $(b)\uparrow \overset{d}{=} \{x \mid b \sqsubseteq x\}$.

Our analysis will rest on the topological properties of continuous dcpos.

Definition 11.6 The **Scott topology** on a dcpo D, written σ_D, consists of all sets U satisfying $(U)\uparrow = U$ and for all directed sets $X \subseteq D$, $\sup X \in U$ implies $X \cap U \neq \emptyset$. In any ω-continuous dcpo, $\{b\Uparrow \mid\ b \in$ basis for $D\}$ is a base for the Scott topology.

The Scott topology does not have the strong separation properties that one may be used to in geometry and analysis. Given a pair of points x and y there will be an open set containing one but not the other. There is no guarantee that there is an open set that plays the reverse role and certainly not that points can be separated by disjoint open sets.

11.1.2 *Valuations and the probabilistic powerdomain*

Valuations play the role of measures, giving quantitative content to open sets but the axioms have to take into account the fact that complements of open sets are not necessarily open so there have to be substitutes for notions like σ-additivity.

The Scott topology allows us to define valuations as continuous functions, on the lattice σ_D, that satisfy modularity.

Definition 11.7 A valuation ν on a dcpo D is a monotone and continuous function from $(\sigma_D, \subseteq)$ to $[0,1]$ that satisfies:

$$(\forall U, V \in \sigma_D \ [\nu(U \cup V) = \nu(U) + \nu(V) - \nu(U \cap V)]$$

The following special valuation plays an important role.

Definition 11.8 For any $x \in D$, the **point valuation** η_x is 1 for all Scott-opens that contain x, and 0 for all others.

This is analogous to the Dirac distribution.

Definition 11.9 The **probabilistic powerdomain** of D, written $\mathcal{P}_{\mathsf{Pr}}(D)$, is the set of all valuations on D ordered by $\nu \sqsubseteq \mu \Leftrightarrow \ (\forall U \in \sigma_D) \ [\nu(U) \leq \mu(U)]$.

If D is ω-continuous, so is $\mathcal{P}_{\mathsf{Pr}}(D)$ with a countable basis given by valuations of the form:

$$r_1 \times \eta_{x1} + \ldots + r_n \times \eta_{xn}$$

where r_i are rationals.

There is a unique extension of valuations to measures on the Borel sets associated with the Scott topology. For ω-continuous dcpos, see [SD80; Jon90; AMESD98; Law82] for the extension theorems.

11.1.3 *The Lawson topology*

The Scott topology meshes closely with the order structure. However, it has weak separation properties and does not capture *by itself* all properties of interest. In particular, Scott-compactness, does not always turn out to be useful.

The closely related Lawson topology is often needed; roughly speaking the Lawson topology gives "negative" information.

Definition 11.10 The **Lawson topology** on a dcpo D, written $\lambda(D)$, is generated by the base $U \setminus (F)\uparrow$, where U is Scott open, and F is a finite subset of D. In any ω-continuous dcpo, a countable base for the Lawson topology is given by $\{b\Uparrow \setminus (\{b_1, \ldots, b_n\})\uparrow \mid\ b, b_i \in$ a basis for $D\}$.

As an example of the Lawson topology consider the domain of streams over a finite alphabet with the prefix ordering. This is an ω-algebraic dcpo. The Lawson topology in this case coincides with the topology induced by the following metric. Given two streams s and t we look for the first position, say the nth, where s and t differ. We then define $d(s,t) := 2^{-n}$. Now the sequence of streams 0^n1^∞ will converge in this metric – hence, in the Lawson topology, to 0^∞. In the Scott topology the sequence of streams 0^n1^∞ is not convergent.

For ω-continuous dcpos D, the sets $(b)\uparrow$ are G_δ in σ_D. Thus, in this case, the Borel algebra generated by $\lambda(D)$ is the same as that generated by σ_D. The Lawson topology on ω-continuous dcpos is separable and metrisable; the Scott topology is T_0 and usually not even T_1. In the case that the Lawson topology is also compact, we get a Polish space for ω-continuous dcpos.

We now explore some consequences of Lawson-compactness. First, the following result (and proof) from [Jun88] relates Scott-compactness and Lawson-closedness.

Lemma 11.1 *In a ω-continuous dcpo, every Scott-compact upper set A is Lawson-closed and is expressible as the countable intersection of Scott-open sets.*

Proof. Consider any Scott-open neighbourhood $O \supseteq A$. Each element $x \in A$ is contained in some basic Scott-open of the form $b\Uparrow, b \in O$. These sets form a cover for A. By Scott-compactness, there is a finite subcover of such sets. Thus, for each $O \supseteq A$, there is a finite collection of b_i such that $A \subseteq \cup_{i=1}^n b_i\Uparrow \subseteq O$. However, $b_i \in O$, so $(b_i)\uparrow \subseteq O$, so we also have

$A \subseteq \cup_{i=1}^{n}(b_i)\uparrow \subseteq O$. For any upper set A, $A = \cap O, O \supseteq A$ with O Scott-open. Hence $A = \cap_O \cup_{i=1}^{n_O} b_i \Uparrow$, and $A = \cap_O \cup_{i=1}^{n_O} (b_i)\uparrow$.

Now $(b_i)\uparrow$ is a (basic) Lawson-closed set, hence A is also Lawson-closed. From ω-continuity, there are only countably many collections of the form $\cup_{i=1}^{n_O} b_i \Uparrow$. □

If D is Lawson-compact, the ordering relation on valuations can be characterised in terms of upper sets. More precisely, the ordering relation between the induced measures on upper sets can be characterised in terms of the ordering on valuations. In the following lemma we will write ν for a valuation and $\hat{\nu}$ for the induced measure on the Borel sets[1]. Thereafter we will revert to using the same symbol for the valuation and its induced measure.

Lemma 11.2 *Let D be Lawson-compact. Let $\nu_1, \nu_2 \in \mathcal{P}_{\mathsf{Pr}}(D)$.*

(1) $\nu_1 \sqsubseteq \nu_2 \Rightarrow$ $(\forall$ *upper Borel* $A).[\hat{\nu_1}(A) \leq \hat{\nu_2}(A)]$
(2) $(\forall$ *Lawson-closed upper* $A)[\hat{\nu_1}(A) \leq \hat{\nu_2}(A)] \Rightarrow \nu_1 \sqsubseteq \nu_2$

Proof.

(1) If A is a Scott-compact upper set, then by Lemma 11.1 it is the countable intersection of Scott-open sets O_i. Let $U_k = \cap_{i=1}^{k} O_i$. Then $(U_k)_{k\in\mathbf{N}}$ is a nested sequence of Scott-open sets decreasing to A. Thus $\hat{\nu_j}(A) = \inf_i \hat{\nu_j}(U_i)$, for $j = 1, 2$ because $\hat{\nu_j}, j = 1, 2$ are measures. But, since the U_i are Scott-opens, we have

$$\hat{\nu_1}(U_i) = \nu_1(U_i) \leq \nu_2(U_i) = \hat{\nu_2}(U_i),$$

and we get – after taking infs – that $\hat{\nu_1}(A) \leq \hat{\nu_2}(A)$, for Scott-compact upper sets A.

For general upper sets A we proceed as follows. Since the domain is a metrisable space, we know [Par67] that $\hat{\nu_j}(A) = \sup\{\hat{\nu_j}(C) \mid C \text{ Lawson-closed}, C \subseteq A\}$ for $j = 1, 2$. Now if C is closed it is Lawson-compact, since it is a closed subset of a Lawson-compact space. Thus, it is also Scott-compact (the Scott topology has fewer sets). Thus its upper set is also Scott-compact. Now we claim that

$$\sup\{\hat{\nu_j}(C) \mid\ C \text{ Lawson-closed}, C \subseteq A\} = \sup\{\hat{\nu_j}(C) \mid C \text{ is a Scott-compact upper set}, C \subseteq A\}.$$

[1] Recall that the Borel sets are the same for the Scott topology and the Lawson topology.

To see this, observe that since every Scott-compact set is Lawson-closed (Lemma 11.1), the right hand side $\leq$ the left hand side. Conversely, for every element in the left hand side, there is a bigger element in the right hand side, namely its upclosure. Hence the two must be equal. But since we know the result for all Scott-compact upper sets, we have our result in the general case.

(2) Let ($\forall$ Lawson-closed upper $C)[\hat{\nu_1}(C) \leq \hat{\nu_2}(C)]$. Since $\lambda(D)$ is metrisable, we have for all Scott-opens $O : \nu_j(O)\hat{\nu_j}(O) = \sup\{\hat{\nu}(C) \mid C \subseteq O,\ C \text{ Lawson-closed}\}, j = 1, 2$. Since D is Lawson-compact, every closed C is also Lawson-compact, and hence Scott-compact. Thus, its upper set is also Scott-compact, and by lemma 11.1 Lawson-closed. Thus, we have: $\nu_j(O) = \sup\{\hat{\nu}(C) \mid C \subseteq O,\ C \text{ upper and Lawson-closed}\}, j = 1, 2$. The result now follows from the assumption on upper Lawson-closed sets.

□

Lawson-compactness is stable under inverse limits [DGJP03].

Lemma 11.3 *Lawson-compact, ω-continuous dcpos are closed under inverse limits.*

Proof. Let $\{(D_i, f_i, g_i)\}$ be an inverse-limit system, i.e.

$$\begin{aligned} f_i &: D_{i+1} \longrightarrow D_i \\ g_{i+1} &: D_i \longrightarrow D_{i+1} \\ f_i \circ g_{i+1} &= 1 \\ g_{i+1} \circ f_i &\sqsubseteq 1 \end{aligned}$$

Here all the f_i, g_i are Scott-continuous, monotone functions. However, one can immediately show that f_i must be Lawson-continuous – one only needs to check that $f^{-1}((x)\uparrow)$ is Lawson-closed, since f^{-1} respects all set operations. However, this is immediate since $f^{-1}((x)\uparrow) = (g(x))\uparrow$, which is Lawson-closed.

D, the inverse limit, is the subspace of $\Pi_i D_i$, given by sequences of the form $\langle d_i \rangle$, where $\forall i. d_i \in D_i, f_i(d_{i+1}) = d_i$. D_i embeds in D via $e_i(x) = \langle d_j \rangle$ where $d_i = x$, $d_j = f_j \circ \ldots \circ f_i(x)$, if $j < i$ and $d_j = g_j \circ \ldots \circ g_{i+1}(x)$ if $j > i$. D has a countable basis given by $\cup_i \{e_i(x_i) \mid x_i \text{ is a basis element of } D_i\}$.

The Lawson topology on the inverse limit is the subspace topology inherited from the product topology of all the domains in the diagram. If all D_is are compact Hausdorff then so is the product, by Tychonoff's theorem. Thus, in order to complete the proof we must show that the sequences which constitute the elements of the bilimit form a closed subset, which would imply that the subspace they form is also compact. Equivalently, we are done if we show that D^c (the complement of D in $\Pi_i D_i$) is open. Note that, $D^c = \cup_i X_i$, where $X_i = (\Pi_{j<i} D_j) \times \{\langle x, y\rangle \mid f_i(y) \neq x\} \times (\Pi_{j>i+1} D_j)$. So, we are done if we prove that $\{\langle x, y\rangle \mid f_i(y) \neq x\}$ is open in $D_i \times D_{i+1}$. Since the space D_i is Hausdorff in the Lawson topology, for every $x \neq f_i(y)$ we have disjoint open sets $A_{xy}, B_{xy} \subseteq D_i$ such that $x \in A_{xy}$, $f_i(y) \in B_{xy}$. Thus, the given set is $\bigcup_{xy} A_{xy} \times f_i^{-1}(B_{xy})$, and thus is open, as f_i is Lawson-continuous. □

11.2 The Domain *Proc*

We fix a (countable) set L of labels and use the Jones-Plotkin probabilistic powerdomain [JP89; Jon90]. For notational convenience, we write $\mathsf{L} \longrightarrow D$ for the product $\prod_{\mathsf{L}} D$ indexed by the set of labels. Processes are given by the recursive domain equation:

$$Proc = \mathsf{L} \longrightarrow \mathcal{P}_{\mathsf{Pr}}(Proc).$$

We will write $\sqsubseteq$ for the partial order in the domain.

Proposition 11.1 *The domain equation*

$$Proc = \mathsf{L} \longrightarrow \mathcal{P}_{\mathsf{Pr}}(Proc)$$

can be solved in the category of ω-continuous, Lawson-compact dcpos.

Proof. By [JT98] the probabilistic powerdomain is closed on Lawson-compact, ω-continuous dcpos. We have already showed closure of ω-continuous, Lawson-compact dcpos under inverse limits. Thus the domain equation can be solved in this category using standard techniques [SP82]. □

This domain can be viewed as a single "universal" labelled Markov process. The next few results show how to define transition probabilities in order to do this. To start with, we have the transition probabilities to Scott-open sets given by the definition of the Jones-Plotkin powerdomain. In Section 11.4 we show how to extend these results to obtain the transition probabilities to arbitrary measurable sets, i.e. the Borel sets generated by the Lawson topology.

Definition 11.11 $\tau_a(p, U) \overset{d}{=} p(a)(U)$ for $p \in \mathit{Proc}$, and U a Scott-open set in *Proc*.

Lemma 11.4 *$\tau_a(., U)$ is upper semicontinuous for each Scott-open U.*

Proof. We need to show that $[\tau_a(., U)]^{-1}(r, 1]$ is a Scott-open set. In order to show that it is upper closed we proceed as follows. Let $p \sqsubseteq q$, and $\tau_a(p, U) > r$. Now we have $\tau_a(q, U) = q(a)(U) \geq p(a)(U) > r$.

In order to show the remaining property of Scott-opens we let $p = \sqcup p_i$, and suppose $\tau_a(p, U) > r$. Now $p(a)(U) = \tau_a(p, U) > r$. The fact that p is the sup of the p_i implies that $\exists i . p_i(a)(U) > r$. □

The domain *Proc* is a ω-continuous domain with a basis given by the following notion of "finite process".

Definition 11.12 A **finite process** is generated by the following grammar:

$$\begin{aligned} q ::= {} & 0 \\ & \mid \{\{a_i \longrightarrow \{(q_{i\,,1}, r_{i\,,1}), \ldots, (q_{i\,,n_i}, r_{i\,,n_i})\}\} \mid i = 1, \ldots, k, \textstyle\sum_{j=1,\ldots,n_i} r_{i,j} \leq 1\} \end{aligned}$$

where a_i are labels and $r_{i,j}$ are real numbers in $[0, 1]$. A finite process is **rational** if all probabilities are rational.

A finite process is interpreted as an element of *Proc* via the following inductive definition.

Definition 11.13

$$\begin{aligned} \llbracket 0 \rrbracket &= \bot, \\ \llbracket \{\{a_i \longrightarrow \{(q_{i\,1}, r_{i\,1}), \ldots, (q_{i\,n_i}, r_{i\,n_i})\}\} \mid i = 1, \ldots, k\} \rrbracket(a) &= \sum_{k=1\ldots n_i} r_{i,n_k} \eta(\llbracket q_{i,n_k} \rrbracket), \ a = a_i \\ &= \bot, \text{ otherwise} \end{aligned}$$

Lemma 11.5 *The set of denotations of finite rational processes is a countable basis for Proc.*

Proof. *Proc* is the limit of the inverse limit system $\{(D_i, f_i, g_i)\}$ where $D_{i+1} = \mathsf{L} \longrightarrow \mathcal{P}_{\mathsf{Pr}}(D_i)$. *Proc* has a countable basis induced by the basis of the D_is. The result now follows from Chapter 5 of [Jon90], which characterises basis elements of the probabilistic powerdomain. □

Definition 11.14 An equivalence relation R on *Proc*, is an **internal bisimulation** if: sRt implies that for all labels a, and all Scott-open R-closed C, $s(a)(C) = t(a)(C)$. We say that s is **internally bisimilar** to t if there exists a internal bisimulation R such that sRt.

Internal bisimulation is an equivalence relation and is the maximum fixed point of the following monotone function F on the lattice of equivalence relations on *Proc* × *Proc*, where the ordering is inclusion :

$$s\,F(R)\,t \text{ if for all labels } a, \text{ and all Scott-open } R\text{-closed } C, s(a)(C) = t(a)(C)$$

Definition 11.15 A preorder R on *Proc* is an **internal simulation** if sRt implies that for all labels a, and all Scott-open R-closed sets C, $s(a)(C) \leq t(a)(C)$.

Internal simulation is a preorder and is the maximum fixed point of the following monotone function G on the lattice of preorders on *Proc* × *Proc*, where the ordering is inclusion :

$$s\,G(R)\,t \text{ if for all labels } a, \text{ and all Scott-open } R\text{-closed } C, s(a)(C) \leq t(a)(C)$$

The following proposition is immediate from the above definition.

Proposition 11.2 $\sqsubseteq$ *of Proc is an internal simulation.*

11.3 $\mathcal{L}_\vee$ as the Logic of *Proc*

In this subsection, we show that $\mathcal{L}_\vee$ is complete for internal simulation. In the context of this subsection, the logic will be interpreted over the domain *Proc*.

Definition 11.16

$$\begin{array}{ll}
p \models \top & \\
p \models \phi_1 \wedge \phi_2 & \text{if } p \models \phi_1 \text{ and } p \models \phi_2 \\
p \models \phi_1 \vee \phi_2 & \text{if } p \models \phi_1 \text{ or } p \models \phi_2 \\
p \models \langle a \rangle_q \phi & \text{if } \exists U : \text{ Scott-open with } U \subseteq [\![\phi]\!] \text{ such that } p(a)(U) > r
\end{array}$$

where $[\![\phi]\!] = \{p \mid p \models \phi\}$.

An important property of the logically definable sets is that they are all Scott-open.

Lemma 11.6 $[\![\phi]\!]$ *is a Scott-open subset of Proc.*

Proof. Recall that a set A is Scott-open if A is up-closed and if whenever a directed sup, say $\sqcup X$, is in A then some member of X is in A. We proceed by induction on the structure of ϕ. Since we have $[\![\top]\!] = \mathit{Proc}$, $[\![\phi_1 \wedge \phi_2]\!] = [\![\phi_1]\!] \cap [\![\phi_2]\!]$ and $[\![\phi_1 \vee \phi_2]\!] = [\![\phi_1]\!] \cup [\![\phi_2]\!]$; it follows immediately that the boolean connectives preserve Scott-openness.

For the case of the modal operator we proceed as follows. We assume that $[\![\phi]\!]$ is Scott-open and show that $[\![\langle a \rangle_q \phi]\!]$ is Scott-open. Let $p \sqsubseteq p'$, then $p(a) \leq p'(a)$, and hence $p(a)([\![\phi]\!]) > q$ implies $p'(a)([\![\phi]\!]) > q$. So, $[\![\langle a \rangle_q \phi]\!]$ is up-closed. Let $p = \sqcup p_i$, and $p \models \langle a \rangle_q \phi$, then $\exists p_i$ such that $p_i(a)[\![\phi]\!] > q$□

Lemma 11.7 *Let p and q be elements of Proc. Then, the following are equivalent:*

(1) $p \sqsubseteq q$*,*
(2) p *is simulated by* q*,*
(3) $p \models \phi$ *implies* $q \models \phi$*.*

Proof. (1) ⇒ (2): From Proposition 11.2 $\sqsubseteq$ is an internal simulation.
(2) ⇒ (3): By structural induction on formulas.
(3) ⇒ (1): First assume that p is a finite process. We will proceed by induction on the height of p. Let $p(a) = (p_1, r_1), \ldots, (p_n, r_n)$. We use Lemma 4.8 of [Jon90]. We need to show that for all up-closed (under $\sqsubseteq$) subsets K of $\{p_1, \ldots, p_n\}$:

$$\sum_{p_i \in K} r_i \leq q(a)(\cup_{p_i \in K} \{x \mid p_i \sqsubseteq x\})$$

First we show that: $\{x \mid p_i \sqsubseteq x\} = \cap \{[\![\psi]\!] \mid p_i \models \psi\}$.

- lhs $\subseteq$ rhs: $p_i \sqsubseteq x \Rightarrow (p_i \models \psi \Rightarrow x \models \psi)$.
- rhs $\subseteq$ lhs: $x \in \cap\{[\![\psi]\!] \mid p_i \models \psi\}$ implies $(p_i \models \psi \Rightarrow x \models \psi)$ implies $p_i \sqsubseteq x$, by the induction hypothesis on p_i.

Next, by distributivity, we see that:

$$\bigcup_{p_i \in K} \bigcap \{[\![\psi]\!] | p_i \models \psi\} = \bigcap_{\langle \psi_1, ..\psi_n \rangle} \{[\![\psi_1 \vee \psi_2 \vee ..\psi_n]\!] \mid p_i \in K, p_i \models \psi_i\}$$

Thus, it suffices to show that for all upclosed (under $\sqsubseteq$) subsets K of $\{p_1, \ldots, p_n\}$:

$$\sum_{p_i \in K} r_i \leq q(a)(\{[\![\psi_1 \vee \psi_2 \vee ..\psi_n]\!] \mid p_i \in K, p_i \models \psi_i\})$$

But, for any formula ϕ, by considering the formulas of the form: $\langle a \rangle_r \phi$, we get $p(a)([\![\phi]\!]) \leq q(a)([\![\phi]\!])$.

Now for general processes we proceed as follows. Let p' be a finite process way-below p, i.e. $p' \ll p$. We will show that $p' \sqsubseteq q$ and hence it will follow (since the finite processes form a basis) that $p \sqsubseteq q$. Every formula satisfied by p' is satisfied by p (since (1) implies (3)) and hence by q (by assumption) and hence by the proof in the preceding paragraph it follows that $p' \sqsubseteq q$. This concludes the proof that (3) implies (1) in the general case. □

Corollary 11.1 *p is internally bisimilar to q, if and only if $p = q$ if and only if $(\forall \phi \ \in \mathcal{L}_\vee)$ $p \models \phi \Leftrightarrow q \models \phi$.*

11.4 Relating *Proc* and LMP

In the preceding sections we have developed the theory of labelled Markov processes from two points of view. First we have the probability-theory view of labelled Markov processes and have developed the notion of finite approximation; second we have the domain theory view based on the Jones-Plotkin powerdomain. Here we show how one can go back and forth between these views: every element of *Proc* is a labelled Makov process and conversely, there is an embedding from labelled Markov processes to *Proc*. This correspondence yields the desired Polish space structure on labelled Markov processes and shows that the simple modal logic we have considered characterises simulation for all labelled Markov processes.

11.4.1 *From Proc to labelled Markov processes*

The main result of this section is that the dcpo *Proc* can be made into a "universal" labelled Markov process, in an appropriate sense.

Let p be an element in the probabilistic powerdomain *Proc*. Consider the putative labelled Markov process $\mathcal{U}_0 = (|Proc|, p, \tau_a)$ where

- we are considering the elements of *Proc* under the Lawson topology (yielding a Polish space) [Law97],
- τ_a is given by the unique extension of the valuation $p(a)(\cdot)$ to measures on the Borel sets associated with the Lawson topology.

With the above notation we are able to state the main theorem of this section; it says that *Proc* defines a labelled Markov process.

Theorem 11.1 *The structure $\mathcal{U}_0 = (|Proc|, p, \tau_a)$ is a labelled Markov process.*

Proof. We only have to prove that $\tau_a(., E)$ is a measurable function for measurable $E \subseteq |Proc|$. We already know this for Scott-open E, by

Lemma 11.4. We now show that this is true for the σ-algebra generated by the Scott topology – since this is the same as the σ-algebra generated by the Lawson topology, this will complete the proof.

$\tau_a(., E^c) = \tau_a(., D) - \tau_a(., E)$, hence if $\tau_a(., E)$ is measurable, so is $\tau_a(., E^c)$.

Let E_i be a countable pairwise disjoint collection of sets. $\tau_a(., \cup_i E_i) = \sum_i \tau_a(., E_i)$, hence if each $\tau_a(., E_i)$, so is $\tau_a(., \cup_i E_i)$. This shows the result for all measurable sets E. □

Now we show that internal simulation – which we know coincides with the domain order from Lemma 11.2 – coincides with simulation on labelled Markov processes. Since we already know that $\mathcal{L}_\vee$ characterises internal simulation we get that $\mathcal{L}_\vee$ gives a logical characterisation of simulation.

Theorem 11.1 *The $\sqsubseteq$ order on Proc is a simulation between the corresponding labelled Markov processes.*

Proof. Now suppose $s \sqsubseteq t$ in *Proc*. Thus for every label a and every Scott-open set U, $s(a)(U) \leq t(a)(U)$. Let R be the relation on $\mathcal{U}_0$ defined by sRt if and only if $s \sqsubseteq t$ in *Proc*. We want to show that R is a simulation. Let X be an R-closed measurable set of $\mathcal{U}_0$. We need to show $\tau_a(s, X) \leq \tau_a(t, X)$. Since X is R-closed in $\mathcal{U}_0$, it is up-closed in *Proc*. The result now follows from lemma 11.2. □

Note how Lawson-compactness – through the use of Lemma 11.2 – played a crucial role here.

11.4.2 *Embedding labelled Markov processes into Proc*

In this subsection we show how one can embed the poset of labelled Markov processes ordered by simulation (henceforth LMP) into the domain *Proc*. This completes the passage between the two views. In order to do this we will embed the finite acyclic labelled Markov processes (FAMPs for short) into *Proc*. Since FAMPs are acyclic they have a well-defined height; we define the embedding function $\psi(.) : FAMP \longrightarrow Proc$ by induction on the height of the DAG:

–$\psi(NIL) = \perp$, for a FAMP of height 0

–Let $(\mathcal{P}, p_0, \rho)$ be a FAMP. Then $\psi(\mathcal{P}) = \psi(p_0)$ is defined by $\psi(p_0)(a_i) = \sum_{p \in P} \rho_{a_i}(p_0, p) * \eta_{\psi(p)}$ where η_p is the point valuation at p.

The next lemma relates satisfaction of logical formulas by FAMPs and by their corresponding elements in the domain. In order to distinguish the two notions of satisfaction we will write $\models_D$ for the domain-theoretic notion and $\models_M$ for the Markov process concept.

Lemma 11.8 *Let $\mathcal{P}$ be a FAMP and ϕ a formula then*

$$\mathcal{P} \models_M \phi \iff \psi(\mathcal{P}) \models_D \phi.$$

Proof. Let $\mathcal{P} = (P, p_0, \rho)$ be a FAMP. We prove by induction on the structure of formulas that for every formula ϕ we have $[\![\phi]\!]_{\mathcal{P}} = P \cap \psi^{-1}([\![\phi]\!]_D)$. The base case and the induction step for conjunction and disjunction are obvious. Assume the claim is true for ϕ, we want to prove it for $\langle a \rangle_q \phi$. By definition of ψ, we have that for every $p \in P$ $\psi(p)(a) = \sum_{p' \in P} \rho_a(p, p')\mu_{\psi(p')}$ (even if p is NIL). Hence by induction hypothesis we have

$$\begin{aligned}\rho_a(p, [\![\phi]\!]_{\mathcal{P}}) &= \rho_a(p, P \cap \psi^{-1}([\![\phi]\!]_D)) \\ &= \sum_{p' \in P} \rho_a(p, p')\mu_{\psi(p')}([\![\phi]\!]_D) \\ &= \psi(p)(a)([\![\phi]\!]_D).\end{aligned}$$

Then we get $[\![\langle a \rangle_q \phi]\!]_{\mathcal{P}} = P \cap \psi^{-1}([\![\langle a \rangle_q \phi]\!]_D)$ which completes the proof that $[\![\phi]\!]_{\mathcal{P}} = P \cap \psi^{-1}([\![\phi]\!]_D)$. Hence we have proved that $\mathcal{P} \models \phi$ if and only if $\psi(\mathcal{P}) \models \phi$ for every formula ϕ. □

It follows from the above lemma that the embedding function is monotone.

Lemma 11.9 *If a FAMP $\mathcal{P}_2$ simulates another FAMP, say $\mathcal{P}_1$, then $\psi(\mathcal{P}_1) \sqsubseteq \psi(\mathcal{P}_2)$.*

Proof. Suppose that $\mathcal{P}_2$ simulates $\mathcal{P}_1$ in the notation of the lemma, and that $\psi(\mathcal{P}_1) \models_D \phi$. Then by lemma 11.8 $\mathcal{P}_1 \models_M \phi$. Since $\mathcal{P}_2$ simulates $\mathcal{P}_1$ we have that $\mathcal{P}_2 \models_M \phi$. Then we have $\psi(\mathcal{P}_2) \models_M \phi$. Thus $\psi(\mathcal{P}_1) \sqsubseteq \psi(\mathcal{P}_2)$, which is what we wanted to prove. □

Now we turn to the task of embedding all labelled Markov processes into *Proc*. Let $\mathcal{S}$ be a labelled Markov process and suppose that $\{\mathcal{P}_i \mid i \in I\}$ is a sequence of FAMPs such that:

(1) $i \leq j \Rightarrow \mathcal{P}_i$ is simulated by $\mathcal{P}_j$,
(2) every formula satisfied by $\mathcal{P}$ is satisfied by some $\mathcal{P}_i$.

In Chapter 9 we have shown how to construct such a family of approximations. Now by Lemma 11.9, condition 1 means that the family

$\{\psi(\mathcal{P}_i) \mid i \in I\}$ is a chain in the domain *Proc*. We define

$$\psi(\mathcal{S}) = \sqcup_{i \in I} \psi(\mathcal{P}_i).$$

From the fact that the sets $[\![\phi]\!]_D$ are open in the Scott topology of *Proc* we know that every formula satisfied by $\psi(\mathcal{S})$ is satisfied by some $\psi(\mathcal{P}_i)$. Together with condition 2 we immediately get that

Proposition 11.3

$$\psi(\mathcal{S}) \models_D \phi \iff \mathcal{S} \models_M \phi.$$

It follows immediately – from this and from the logical characterisation of bisimulation – that

Corollary 11.2 *If* $\psi(\mathcal{S}_1) = \psi(\mathcal{S}_2)$ *then* $\mathcal{S}_1$ *is bisimilar to* $\mathcal{S}_2$.

What we have done is to show that the recursive domain equation defines a very special kind of *universal* LMP. In categorical jargon we have constructed a *final object*. This gives co-inductive techniques for reasoning about bisimulation of LMPs.

Chapter 12

Real-Time and Continuous Stochastic Logic

So far we have considered the LMP model where the state space may be continuous but the transitions are discrete jumps in time triggered by external actions or labels. Real-time systems are extremely important in practice, particularly in performance evaluation and verification of real-time systems. Logics and model checking techniques for them are flourishing; see, for example, the excellent recent book by Baier and Katoen [BK08]. In this chapter we consider real-time systems with, possibly, continuous state spaces and prove a logical characterisation of bisimulation in this context.

In the systems that we consider, transitions are still instantaneous but there may be a random delay in each state before the transition is taken. Thus, the system still has a discrete character. Further, we assume that the delay times are exponentially distributed so that the system in a given state has no "memory" of how long it has been there. This is an additional Markovian assumption; such systems are called continuous-time Markov chains or CTMCs for short. Logics and bisimulation techniques for systems where the dynamics is a random flow governed by a stochastic differential equation is still open.

A logic for CTMCs called *Continuous Stochastic Logic* (henceforth **CSL**) was introduced by Aziz *et al.* [ASSB96] and a few years later Baier, Haverkort, Hermanns and Katoen [BHHK00] came up with an ingenious model checking algorithm for **CSL**. Along the way, they showed that if two CTMCs were *bisimilar* then they would satisfy exactly the same formulas of **CSL**. This gives the basic sanity check that the logic is not overly sensitive. In response to a question of Joost-Pieter Katoen to the author, Desharnais and the author [DP03] showed that in fact a small subset of **CSL** suffices to characterise bisimulation for CTMCs, even when the state space is a continuum. In this chapter we prove this result. The basic intuition behind the proof is that **CSL** is rich enough to pin down the exact transition rates. So by looking at the right formulas we can force the systems to have matching transition rates.

In the course of the proof it is necessary to prove that the set of so-called "Zeno" paths has measure 0. This result has been proved for discrete systems [dA97; Bai] and it easily extends to continuous systems. In the paper [DP03] we gave an unnecessarily sophisticated proof using Fredholm operators, but this, while quite amusing, is not necessary.

The next section, Section 12.1 gives basic background on CTMCs extended to continuous state systems. The following section discusses the space paths for a CTMC. Section 12.3 following gives a summary of the logic **CSL** and its semantics. In Section 12.4 we prove a general theorem that is useful for establishing logical characterisation results. The logical characterisation theorem of bisimulation for CTMCs and **CSL** is proved in Section 12.5.

12.1 Background

In this section we generalise the CTMCs and the logic **CSL** to processes with a continuous state space. For simplicity, we follow [BHHK00] and do not consider processes with labelled transitions. The results, however, are not affected – except in trivial notational ways – if we add labels. As in the rest of the book we assume that the state spaces of the CTMCs are analytic and we use the same machinery that we used to prove the logical characterisation theorem for LMPs.

Let AP be a fixed, finite set of atomic propositions.

Definition 12.1 A **continuous-time Markov process** or CTMP is a tuple (S, Σ, R, L) where (S, Σ) forms an analytic space, $R : S \times \Sigma \longrightarrow \mathbf{R}_{\geq 0}$ is a rate function: it is measurable in its first coordinate and a measure on its second coordinate; and $L : S \longrightarrow 2^{AP}$ is the labelling (measurable) function which assigns to each state $s \in S$ the set $L(s)$ of atomic propositions of AP that are valid in s. One may assume an initial distribution α on the state space.

We write $E(s)$ for $R(s, S)$ (the exit rate out of s). The set of absorbing states $R^{-1}\{0\}$ is written Abs and $Can := S \setminus Abs$ is the set of states that can make a transition. We also write $P(s, X) = R(s, X)/E(s)$ if $s \in Can$: it is the probability of jumping to a state in X, given that we start in the state s. If $s \in Abs$, $P(s, X) = 0$. Note that $P(\cdot, X)$ is measurable since it is a quotient of measurable functions (see [Bil95] for example).

Given a state s and a (measurable) set of states say X, the transition rate – transition probability per unit time – for jumping from s to a state in X – is $R(s, X)$. If there are two disjoint sets of states that s can jump to we will have competing transitions and there will be a *race condition* between these transitions. If at time 0 the system is known to be in state s then at time t the probability that the transition to a state in X has been triggered is

$$1 - e^{-R(s,X)\cdot t}.$$

One cannot say that this is the probability that the system will be in a state of X at time t. The probability that at time t the system reaches a state in X in one transition is 0 if $s \in Abs$, and

$$\frac{R(s, X)}{E(s)}[1 - e^{-E(s)\cdot t}]$$

otherwise. As a last example, the probability of leaving s within the interval of time $[t, u]$ is

$$e^{-E(s)t}(1 - e^{-E(s)(u-t)}) = e^{-E(s)t} - e^{-E(s)u},$$

because the process must stay in s for t units of time and then leave s within $u - t$ units of time. Note that this yields 0 if $s \in Abs$.

12.2 Spaces of Paths in a CTMP

In order to understand the behaviour of a CTMP one needs to understand the executions. Thus the set of possible paths and the measures on them play a significant role. One important aspect of continuous time systems is that each transition takes some time; the number of steps by itself is not the correct way of determining elapsed time. In such a case it is possible to have a so-called Zeno path, that is, a path with infinitely many steps but in which the elapsed time intervals decrease fast enough that the total elapsed time is finite. The presence of such paths greatly complicates any calculations. Often – in semantics – one rules out such paths by fiat, usually by invoking some kind of fairness property. In probabilistic treatments the information about transition probabilities is supposed to be a substitute for fairness, so we cannot just rule out Zeno paths. What can be shown instead is that the probability of Zeno paths is zero *provided that the transition rates are bounded.*

It makes sense on physical grounds to insist that the transition rates be bounded. Clearly in finite state systems the transition rates will have some maximum value. In infinite state systems if we allow the rates to grow without bound then the probability of Zeno paths is no longer zero. An explicit example due to Christel Baier shows that this is the case. It is easy to reconstruct the example by considering a countable state space and making the transition rates increase exponentially. In such systems the transition rates grow above any prescribed limit. This is as unphysical as allowing speeds that exceed any limit! The expected time in a state is the inverse of the transition rate out of that state and clearly once one reaches times like 10^{-33} seconds one has an unphysical situation. This is the time unit that can be constructed from the fundamental physical constants G, h, c and it is widely accepted that our usual notions of space and times break down. In the realm of macroscopic objects that usually concern computer science and engineering these time scales are absurd.

Definition 12.2 A *path* is a finite or infinite sequence $s_0, t_0, s_1, t_1, \ldots$ where $t_i \in \mathbf{R}^+$ for $i \in \mathbf{N}$ and $s_i \in \mathit{Can}$ for all $i \in \mathbf{N}$ except for the last state s_l if the sequence is finite in which case $s_l \in \mathit{Abs}$. Let σ be an infinite path; we use the following notation[1]:

$\sigma[i] = s_i$	the i-th state of σ
$\delta(\sigma, i) = t_i$	the time spent in s_i
$\sigma@t = \sigma[i]$	the state of σ at time t i is the least index such that $t < \sum_0^i t_j$
Path	the set of all paths
$\mathit{Path}(s)$	the set of paths starting in s

If σ is finite and ends in s_l, $\sigma[i]$ and $\delta(\sigma, i)$ are defined as above for $i \leq l$, whereas $\delta(\sigma, i) = \infty$, and $\sigma@t = s_l$ for $t \geq \sum_0^{l-1} t_j$.

With $\leq$ in the definition one may be in a situation where no value of t can satisfy the formula. Indeed, we need that for some t in the right interval $\sigma@t = s_i \models \phi'$ for some i and for all $t' = t - \epsilon$, $\sigma@t' = s_{i-1} \models \phi$. As an example, consider the question what is $\sigma@t$ for $t_0 \leq t \leq t_0 + t_1$, strictly in between returns to state s_1? Thus if s_1 is the state at which ϕ' is satisfied (and ϕ is not), we must find a time t such that σ at all time less

[1] The definition of $\sigma@t$ differs from that in [BHHK00] where the inequality is not strict. It is a technical point to make the temporal Until operator of **CSL** satisfiable.

than t returns s_0, this forces us to define $\sigma@t_0$ to be s_1. Hence the strict inequality is necessary here.

The Borel space of paths $\mathcal{F}(Path)$ is generated by sets of paths of the form

- $X_0 \times I_0 \times X_1 \times I_1 \times \ldots I_{n-1} \times X_n \times (\mathbf{R} \times S)^\infty$,
- $X_0 \times I_0 \times X_1 \times I_1 \times \ldots \times X_{n-1} \times I_{n-1} \times Y$

with $X_i \in \Sigma$, I_i an interval of the reals with rational bounds and $n \in \mathbf{N}$ and $Y \in (\Sigma \cap Abs)$. It is not hard to prove that these sets form a semi-ring. This semi-ring, which we denote by $SR(Path)$, is countable because the intervals involved have rational bounds. We often write the tail $(\mathbf{R} \times S)^\infty$ simply as ∞ to simplify the notation.

Given a distribution α on the set of states, we define a probability measure Pr_α on paths as follows. Let $X_0 \times I_0 \times X_1 \times I_1 \times \ldots I_{n-1} \times X_n \times \infty$ be a set (of infinite paths) in $SR(Path)$ and let us write $e(i)$ for the probability $e^{-E(x_i)t_i} - e^{-E(x_i)u_i}$ of leaving x_i within the interval of time $I_i = [t_i, u_i]$. Then

$$\begin{aligned}
&Pr_\alpha(X_0 \times I_0 \times X_1 \times I_1 \times \ldots \times I_{n-1} \times X_n \times \infty) = \\
&\qquad \int_{X_0} e(0) \int_{X_1} e(1) \ldots \int_{X_{n-1}} (e(n-1) P(x_{n-1}, X_n)) \\
&\qquad\qquad P(x_{n-2}, dx_{n-1}) \ldots P(x_0, dx_1) \alpha(dx_0).
\end{aligned}$$

The measure is defined in the same way for a set in $SR(Path)$ of finite type by replacing X_n with Y.

This set function is easily shown to be countably additive on the semi-ring $SR(Path)$ and hence has a unique extension to a measure on $\mathcal{F}(Path)$.

We write $Pr_s(X)$ if we consider the state s as a starting point and hence its initial distribution.

We allow absorbing states in infinite paths because the set of infinite sequences containing absorbing states is of measure 0. In general, it is not necessary to distinguish between the two cases of paths, i.e. those involving absorbing states and those that do not because the measure handles absorbing states. In manipulating sets of paths, we usually decompose them according to the number of transitions in the paths. This way, absorbing states are taken into account. On the other hand, if we want a set of paths of finite type to get a meaningful value (i.e. > 0) we must use a multiple integral with the right number of integrations, that is, the number of integrations must be the number of transitions of the paths in the set.

The following theorem states that Zeno paths have measure 0.

Theorem 12.1 *The set of paths having a sequence of time that converges is of measure* 0 *provided that the rates are bounded; i.e.* $\sup_{s \in S} R(s, S)$ *exists.*

A proof is given in [DP03].

12.3 The Logic CSL

In this section, we recall the logic introduced in [BHHK00] and its semantics. The version of **CSL** [ASSB96] originally introduced by Aziz *et al.* does not have the next-state formula, it has the *Until* construct which allows one to express most path modalities. For our purposes the next-state formula is very useful and is standard in such logics. The version of Baier *et al.* is essentially like ours, except that they also talk about rewards. In [BHHK00], the logic **CSL** was augmented with a very important formula $S_{\bowtie p}(\phi)$ to represent steady-state properties. We do not need it for the purposes of the logical characterisation proof so we omit it.

Definition 12.3 Our version of the logic **CSL** [ASSB96] has the following syntax. Let $a \in AP$, $p \in [0, 1] \cap Q$ and $\bowtie \in \{<, \leq, \geq, >\}$. State formulas ϕ are defined by

$$\phi := \mathsf{T} \mid a \mid \neg\phi \mid \phi \wedge \phi' \mid P_{\bowtie p}(\psi)$$

where ψ is a path formula constructed by

$$\psi := X\phi \mid X^{[t,u]}\phi \mid \phi U \phi' \mid \phi U^{[t,u]}\phi';$$

for t, u rational.

Note that there are countably many formulas since AP is countable and p, t, u are rationals.

Meaning of formulas

Given a CTMP $\mathcal{S} = (S, \Sigma, R, L)$ and $a \in AP$, the definition of the satisfaction relation $\models$ over state formulas is given by induction:

$$\begin{array}{ll}
s \models \mathsf{T} & \text{for all } s \in S; \\
s \models a & \text{iff } a \in L(s); \\
s \models \phi \wedge \phi' & \text{iff } s \models \phi \text{ and } s \models \phi'; \\
s \models \neg\phi & \text{iff } s \not\models \phi; \\
s \models P_{\bowtie p}\psi & \text{iff } Pr_s\{\sigma \in Path(s) : \sigma \models \psi\} \bowtie p.
\end{array}$$

Note that the set $\{\sigma \in Path(s) : \sigma \models \psi\}$ is measurable; this will be shown in the next lemma below. To be more formal, we could have considered at this point the greatest measurable set contained in it. The semantics of path formulas is defined as follows.

$$\begin{array}{ll}
\sigma \models X\phi & \text{iff } \sigma[1] \text{ is defined and } \sigma[1] \models \phi; \\
\sigma \models X^{[t,u]}\phi & \text{iff } \sigma \models X\phi \text{ and } \delta(\sigma, 0) = t_0 \in [t, u]; \\
\sigma \models \phi U \phi' & \text{iff } \exists k \geq 0 \,.\, \sigma[k] \models \phi' \text{ and } \forall 0 \leq i < k, \sigma[i] \models \phi \\
\sigma \models \phi U^{[t,u]} \phi' & \text{iff } \exists t^* \in [t, u]. \sigma@t^* \models \phi' \text{ and } \forall 0 \leq t' < t^*, \sigma@t' \models \phi.
\end{array}$$

We write $[\![\phi]\!]_{\mathcal{S}}$ for the set $\{s \in S \mid s \models \phi\}$. We often omit the subscript when no confusion can arise. Similarly, we write $[\![\psi]\!]_s = \{\sigma \in Path(s) : \sigma \models \psi\}$. The following lemma shows that these sets are measurable in $\mathcal{S}$.

Note that the logic **CSL** could be extended to witness *labelled* transitions, by replacing X with $\langle a \rangle$. In this case, the formula $\langle a \rangle_p \phi$ from [DGJP02a] would be represented in **CSL** by the formula $P_{>p}(\langle a \rangle \phi)$.

Lemma 12.1 *Let $\mathcal{S} = (S, \Sigma, R, L)$ be a CTMP. Then*

*(1) for every formula ϕ of **CSL**, $[\![\phi]\!] \in \Sigma$;*
*(2) for every path formula ψ of **CSL** and every $s \in S$, $[\![\psi]\!]_s \in \mathcal{F}(Path)$.*

Proof. We prove the two statements in parallel using structural induction. Trivially $[\![\mathsf{T}]\!] = S \in \Sigma$. For every $a \in AP$, $[\![a]\!] \in \Sigma$ because L is measurable. Conjunction and negation are trivial since finite intersections and complements of measurable sets are measurable. To prove that $[\![P_{\bowtie p}(\psi)]\!]$ is measurable for every path formula ψ, we will prove that the function $Pr_{(\cdot)}([\![\psi]\!]_{(\cdot)}) : S \longrightarrow [0, 1]$ is measurable.

Now assume $[\![\phi]\!] \in \Sigma$. Consider the path formula $X^{[t,u]}\phi$ and fix $s \in S$. We have

$$[\![X^{[t,u]}\phi]\!]_s = \{s, t_0, s_1, \cdots : t_0 \in [t, u], s_1 \models \phi\} = \{s\} \times [t, u] \times [\![\phi]\!] \times \infty$$

and hence $[\![X^{[t,u]}\phi]\!]_s \in \mathcal{F}(Path)$. Now $Pr_{(\cdot)}([\![X^{[t,u]}\phi]\!]_{(\cdot)}) : S \longrightarrow [0,1]$ satisfies

$$Pr_s([\![X^{[t,u]}\phi]\!]_s) = (e^{-E(s)t} - e^{-E(s)u})\frac{R(s,[\![\phi]\!])}{E(s)}.$$

Now products, differences and compositions of measurable functions are measurable by standard results in measure theory [Ash72; Bil95; Rud66; Hal50]. Thus our function above is measurable in s since it is constructed as a combination (product, difference, quotient, composition) of the measurable functions $E(s)$ and $R(s,[\![\phi]\!])$. Consequently, $\{s : Pr_s([\![X^{[t,u]}\phi]\!]_s) \bowtie p\} = [\![P_{\bowtie p}(X^{[t,u]}\phi)]\!] \in \Sigma$ because it is the inverse image under a measurable function of a measurable subset of $[0,1]$, namely the subset defined by $\bowtie p$.

We now prove that the path formula $\phi U^{[t,u]}\phi'$ describes a measurable set of paths. In the following we write out the set of paths satisfying the Until formula as a union, indexed by the number of transitions it makes before satisfying ϕ'. Each such union is itself written as the union over the possible *rational* times at which the transitions occur. It suffices to use intervals with rational end points since every possible transition time will be represented by some interval. Note that we are not claiming that we have a disjoint union. We specify in each line the time at which the paths satisfy the second condition of the Until formula and use the notation $\sigma = s, t_0, s_1, t_1, \ldots$. Note that if s does not satisfy ϕ then it cannot satisfy the Until formula so we assume that $s \models \phi$ in the following.

$$\begin{aligned}
&[\![\phi U^{[t,u]}\phi']\!]_s \\
&= \cup_i\{\sigma \in Path(s) : \exists T \in [t,u]\, \sigma@t{=}s_i \models \phi' \text{ and } \forall t' < T, \sigma@t' \models \phi\} \\
&= (\{s\} \cap [\![\phi']\!]) \times (t,\infty) \times \infty \qquad s \models \phi' \text{ and } T \text{ in } [t, \min\{t_0,u\}) \\
&\quad \cup\ \{s\} \times [t,u] \times [\![\phi']\!] \times \infty \qquad s_1 \models \phi' \text{ and } T = t_0 \\
&\quad \cup \bigcup_{u_1<u_2<t} \{s\} \times [u_1,u_2] \times [\![\phi \wedge \phi']\!] \times (t-u_1,\infty) \times \infty \\
&\qquad\qquad\qquad s_1 \models \phi \wedge \phi' \text{ and } T \in [t, \min\{t_0+t_1,u\}) \\
&\quad \cup \bigcup_{\substack{u_1<u_2<u \\ u_2-u_1<u-t}} \{s\} \times [u_1,u_2] \times [\![\phi]\!] \times [t-u_1, u-u_2] \times [\![\phi']\!] \times \infty \\
&\qquad\qquad\qquad s_2 \models \phi' \text{ and } T = t_0+t_1 \\
&\quad \cup \bigcup_{\substack{u_1<u_2 \\ u_3<u_4 \\ u_2+u_4<t}} \{s\} \times [u_1,u_2] \times [\![\phi]\!] \times [u_3,u_4] \times [\![\phi \wedge \phi']\!] \times (t-u_1-u_3,\infty) \times \infty \\
&\qquad\qquad\qquad s_2 \models \phi \wedge \phi' \text{ and } T \in [t, \min\{t_0+t_1+t_2,u\}). \\
&\quad \vdots
\end{aligned}$$

In the expression following the last equality sign above, the first line is for the case where s already satisfies ϕ' and the system stays in this state until at least t. In the second line we have the situation where at time t_0 the system jumps to s_1 which satisfies ϕ'. In the third line we have the case where the jump to s_1 occurs between u_1 and u_2 but before t. In this case we must have s_1 satisfying both ϕ and ϕ' and staying in this state until time t. The subsequent lines follow the same pattern.

Since the set in question is expressed as a countable union of measurable sets, it is measurable.

Now we have to show that $Pr_s([\![\phi U^{[t,u]}\phi']\!]_s)$ viewed as a function of s is a measurable function. We have expressed the set $[\![\phi U^{[t,u]}\phi']\!]_s$ as a countable but not disjoint union. Let us consider what the intersections of two of these sets can look like. If the number of jumps before ϕ' is satisfied is different in the two sets then the intersection will be empty. In other cases, if we intersect two sets from the union occurring on the same line, we will be intersecting two sets where the time intervals overlap. In this case the intersection will be in the form of a basic measurable set of paths. If we apply the $Pr_s(\cdot)$ to such a set we clearly get a measurable function of s. Thus when we form finite intersections we get measurable functions of s. Now when we have a finite union - say of the family $U_1, U_2, \dots U_n$ - we can write the probability as[2]

$$\sum_{i=1}^{n} Pr_s(U_i) - \sum_{i\neq j} Pr_s(U_i \cap U_j) + - \dots$$

But each such term defines a measurable function of s so the combination, being constructed from sums and products is also measurable. When we have a countable union we can construct the function $Pr_s(\cdot)$ as the sup of the functions for the finite unions. Since the sup of a family of measurable functions is a measurable function we are done. □

12.4 A General Technique for Relating Bisimulation and Logic

Logical characterisation proofs arise in several closely related places. These proofs tend to be somewhat similar. In this section, we formulate – in a more general and useful form – the theorems that were used implicitly in

[2] This purely combinatorial fact is called the "principle of inclusion-exclusion".

all these proofs. This extracts the essence of the logical characterisation proof of Chapter 7.

Before we give this general form we need some notation. Let $\mathcal{F}$ be a family of measurable sets, i.e. $\mathcal{F} \subseteq \Sigma$. There is an induced equivalence relation, written $\equiv_{\mathcal{F}}$, defined by

$$x \equiv_{\mathcal{F}} y \leftrightarrow \forall A \in \mathcal{F}.(x \in A \leftrightarrow y \in A).$$

In other words two states are equivalent if they belong to exactly the same sets of $\mathcal{F}$. We write $cl(\mathcal{F})$ for the collection of measurable sets closed under the equivalence relation $\equiv_{\mathcal{F}}$ (i.e. measurable unions of equivalence classes). The general theorem can now be stated.

Theorem 12.2 *Let (S, Σ) be an analytic space. Let $\mathcal{F} \subseteq \Sigma$ be countable and closed under intersection and let $S \in \mathcal{F}$. Then if two measures agree on $\mathcal{F}$ then they agree on $cl(\mathcal{F})$.*

It is worth summarising how to use this result. Typically, the set $\mathcal{F}$ will contain the meaning of basic formulas and the measures will be the transition probabilities from two equivalent states. Consequently, to prove that logical equivalence implies bisimilarity, we only have to prove that two logically equivalent states have the same value on transitions to the set of states that satisfy some formula – the definable sets of the logic. If the logic is indeed rich enough to characterise bisimulation then the transition probabilities to the definable sets will force the transition probabilities to agree on all sets and thus be bisimilar. We see that we need our logic to have conjunction and to have only countably many formulas – both are very mild restrictions – and to be rich enough to encode the transition probabilities (or rates) to definable sets. If the logic has these basic properties then a completeness proof can now be produced routinely. Our main Theorem 12.3 is precisely such an application of this result.

One major step towards proving Theorem 12.2 is given by the λ-π theorem. We recall it here, it is Proposition 2.10.

Proposition 12.1 *Let X be a set and $\mathcal{A}$ a family of subsets of X, closed under finite intersections, and such that X is a countable union of sets in $\mathcal{A}$. Let $\sigma(\mathcal{A})$ be the σ-field generated by $\mathcal{A}$. Suppose that μ_1, μ_2 are finite measures on $\sigma(\mathcal{A})$. If they agree on $\mathcal{A}$ then they agree on $\sigma(\mathcal{A})$.*

One can see that the two theorems are very similar and that the only result that we need to prove in order to get Theorem 12.2 from the λ-π theorem is that $\sigma(\mathcal{F}) = cl(\mathcal{F})$. This is exactly the statement of Lemma 12.4. This lemma requires the same results that we used in Chapter 7, namely the unique structure theorem, which we had in Chapter 7 as Lemma 7.2.

Lemma 12.2 *Let (S, Σ) be an analytic space and let Σ_0 be a countably generated sub-σ-field of Σ that separates points in S. Then $\Sigma_0 = \Sigma$.*

Recall that a σ-field separates points if every pair of points is separated by a set in the σ-field, that is, there is some set in the σ-field that contains only one of these points. The following lemma is Lemma 7.3 from Chapter 7.

Lemma 12.3 *Let (S, Σ) be an analytic space and let $\sim$ be an equivalence relation on S. Assume that there is a sequence $f_1, f_2, \ldots, f_n, \ldots$ of measurable real-valued functions on S such that for any pair of points x, y in S one has $x \sim y$ if and only if $\forall n. f_n(x) = f_n(y)$. Then $S/\sim$ is also an analytic space and the trivial quotient map $q : S \longrightarrow S/\sim$ is measurable.*

Now with these two properties of analytic spaces in hand we can prove the following key lemma.

Lemma 12.4 *Let (S, Σ) be an analytic space. Let $\mathcal{F} \subseteq \Sigma$ be countable and assume $S \in \mathcal{F}$. Then $cl(\mathcal{F}) = \sigma(\mathcal{F})$.*

The proof is exactly as in Chapter 7 and is omitted. This lemma establishes Theorem 12.2 immediately.

12.5 Bisimilarity and CSL

In this section, we introduce the definition of bisimulation for CTMPs and prove that it coincides with the equivalence induced by the logic. We prove this result by applying Theorem 12.2 from the preceding section.

Notation Let F be a set of formulas of **CSL**; then $L_F(s)$ is the set of formulas of F that are satisfied by s. Let $\equiv$ be an equivalence relation on S; then $cl(\equiv)$ contains all the closed sets w.r.t. $\equiv$, that is,

$$cl(\equiv) = \{\sigma \in \Sigma : \text{ if } s \in \sigma \text{ and } s \equiv s' \text{ then } s' \in \sigma\}.$$

The following definition of bisimulation generalises the definition of F-bisimulation for discrete CTMCs introduced in [BHHK00].

Definition 12.4 Let F be a set of formulas. An equivalence relation $\equiv$ is an F-bisimulation if whenever $s \equiv s'$, we have $L_F(s) = L_F(s')$ and for every $C \in cl(\equiv)$, $R(s, C) = R(s', C)$.

F is intended to be the set of observable formulas. If we take $F = AP$ then F-bisimulation is standard bisimulation. If we take $F = \{\mathsf{T}\}$ and hence ignore atomic propositions, we get the analogue of probabilistic bisimulation as defined by Larsen and Skou for discrete probabilistic processes (with only one label on the transitions). This parametrisation allows us a more flexible treatment of bisimulation.

Obviously, satisfying only formulas in F would be a far too weak condition for two states to be bisimilar: for example, if $F = \{\mathsf{T}\}$. Bisimulation involves matching transitions and it is necessary that the set of formulas that defines the equivalence between states be closed under the constructor $P_{\bowtie p}(X^{[0,t]}(\cdot))$. One must keep in mind that F is just a restriction on bisimulation that we can impose with the help of atomic propositions. The bigger F is, the fewer states are going to be bisimilar.

Definition 12.5 If F is a set of **CSL** formulas then the closure of F under conjunction $\wedge$ and the operator $P_{\bowtie p}(X^{[0,t]}(\cdot))$ is written $\overline{F}$.

Theorem 12.3 *Let F be a set of formulas that contains the trivial formula T. If two states of a CTMP satisfy the same formulas of $\overline{F}$ then they are F-bisimilar.*

Proof. Let F be a set of formulas; then $\mathcal{F} = \{[\![\phi]\!] : \phi \in \overline{F}\}$ is countable and closed under intersection. We write $s \equiv s'$ if s and s' satisfy the same formulas of $\overline{F}$. We show that $\equiv$ is an F-bisimulation. Let $s \equiv u$; then $L_F(s) = L_F(u)$ trivially. We want to prove that $R(s, C) = R(u, C)$ for every $C \in cl(\equiv)$. By Theorem 12.2, we only have to prove that it is true for $C \in \mathcal{F}$. We first prove that $E(s) = E(u)$. Since $s \equiv u$, s and u satisfy the same formulas of the form $P_{\bowtie p}(X^{[0,t]}\phi)$ where ϕ is a state formula constructed from formulas in $\overline{F}$; consequently, we have

$$Pr_s(\{\sigma \in Path_s : \sigma \models X^{[0,t]}\phi\}) = P_u(\{\sigma \in Path_u : \sigma \models X^{[0,t]}\phi\}). \quad (12.1)$$

Consider the case where $\phi = \mathsf{T} \in \overline{F}$. Then

$$1 - e^{-E(s)t} = 1 - e^{-E(u)t}$$

which implies that $E(s) = E(u)$.

We now prove that $R(s, [\![\phi]\!]) = R(u, [\![\phi]\!])$ for every formula $\phi \in F$. We get from Equation 12.1 that

$$Pr_s(\{\sigma \in Path_s : \sigma \models X^{[0,t]}\phi\}) = (1 - e^{-E(s)t})\frac{R(s, [\![\phi]\!])}{E(s)}.$$

Then

$$(1 - e^{-E(s)t})\frac{R(s, [\![\phi]\!])}{E(s)} = (1 - e^{-E(u)t})\frac{R(u, [\![\phi]\!])}{E(u)}$$

which implies that $R(s, [\![\phi]\!]) = R(u, [\![\phi]\!])$.

By Theorem 12.2, we have that $R(s, A) = R(u, A)$ for every $A \in cl(\equiv)$ as wanted. □

Not all the operators of the logic are needed in the preceding proof. In particular, there is no use of constant, negation and no use of Until. We have seen this phenomenon before with the logical characterisation of LMPs, namely that one does not need very many formulas to get a logic rich enough to get a characterisation of bisimulation.

Let $\mathbf{CSL}_F$ denote the smallest set of formulas of **CSL** containing F and closed under **CSL** operators. The following theorem has been proven for CTMCs in [BHHK00].

Theorem 12.4 *If two states are F-bisimilar, then they satisfy the same formulas of $\boldsymbol{CSL}_F$.*

Proof. The proof is an easy induction on formulas. The strategy is to show that if $\equiv$ is an F-bisimulation, then $[\![\phi]\!]$ is in $cl(\equiv)$, for every state-formula ϕ. If $s \equiv s'$ and $[\![\phi]\!] \in cl(\equiv)$ by induction, then definition of bisimulation implies that $R(s, [\![\phi]\!]) = R(s', [\![\phi]\!])$ and also that $R(s, S) = R(s', S)$ since S is certainly in $cl(\equiv)$. Then we use the equations developed in Lemma 12.1 for $[\![X^{[t,u]}\phi]\!]_s$ and $[\![\phi U^{[t,u]}\phi']\!]_s$. Measurable functions representing the probabilities to these sets are in terms of $R(s, [\![\phi]\!])$ and $R(s, S)$; manipulations of measurable functions and sets complete the proof. □

What happens if the delays are not memoryless? In that case one cannot even define a bisimulation relation on states because the time spent in the state becomes important. One approach developed by Gupta *et al.* [GJP04; GJP06] is to consider *uniformities* which lie between metrics and relations.

This page intentionally left blank

Chapter 13

Related Work

In this short book the focus has been on LMPs with all the nondeterminism associated with the choice of label or action. There are a number of directions that have not been explored; here I give a brief summary of other directions with some pointers to the literature. Many of these are areas of active research and the references to the literature cannot be complete.

13.1 Mathematical Foundations

The key technical result from which most of this work has flowed is the logical characterisation of bisimulation. In this book the proof was based on a definition of bisimulation that mimicked as closely as possible the definition of probabilistic bisimulation due to Larsen and Skou [LS91].

In the original treatment of the subject [DEP02] the definition of probabilistic bisimulation was based on the concept of "spans of zigzags". The idea is to define a *functional* analogue of bisimulation called a zigzag morphism. We recall the definition here

Definition 13.1 A function f from (S, Σ, τ_a) to (S', Σ', τ'_a) is *a zigzag* if it satisfies the properties:

(1) f is surjective;
(2) f is measurable;
(3) $\forall a \in \mathcal{A}, s \in S, \sigma' \in \Sigma', \;\; \tau_a(s, f^{-1}(\sigma')) = \tau'_a(f(s), \sigma')$.

A bisimulation between two LMPs is then a span of such zigzags as shown in the following diagram:

$$\begin{tikzcd}
& (S,\tau) \arrow[dl, "f"'] \arrow[dr, "g"] & \\
(S_1,\tau_1) & & (S_2,\tau_2)
\end{tikzcd}$$

Here the bisimulation relation is represented by the two zigzags f and g. In [DEP02] it is shown that for discrete systems this coincides with the Larsen and Skou definition, however, the proof requires some calculation. For the case of continuous-state systems it requires some work to show that this definition yields an equivalence relation.

The difficult part is showing that one has transitivity. In order to do this one starts with the diagram below

$$\begin{tikzcd}
& (S_4,\tau_4) \arrow[dl] \arrow[dr] & & (S_5,\tau_5) \arrow[dl] \arrow[dr] & \\
(S_1,\tau_1) & & (S_2,\tau_2) & & (S_3,\tau_3)
\end{tikzcd}$$

and looks for a construction that will allow one to complete the upper diamond to produce the figure below:

$$\begin{tikzcd}
& & (S_6,\tau_6) \arrow[dl] \arrow[dr] & & \\
& (S_4,\tau_4) \arrow[dl] \arrow[dr] & & (S_5,\tau_5) \arrow[dl] \arrow[dr] & \\
(S_1,\tau_1) & & (S_2,\tau_2) & & (S_3,\tau_3)
\end{tikzcd}$$

Unfortunately it is very hard to find a construction that takes the diagram

$$\begin{tikzcd}
(S_4,\tau_4) \arrow[dr] & & (S_5,\tau_5) \arrow[dl] \\
& (S_2,\tau_2) &
\end{tikzcd}$$

and completes it to produce the diagram

$$\begin{tikzcd}
& (S_6, \tau_6) \arrow[dl] \arrow[dr] & \\
(S_4, \tau_4) \arrow[dr] & & (S_5, \tau_5) \arrow[dl] \\
& (S_2, \tau_2) &
\end{tikzcd}$$

If one has pullbacks in the appropriate category one can, of course, complete this picture always. If one does not have pullbacks one can make do with *weak pullbacks*. These are like pullbacks, one can always complete the square but the mediating morphism required by universality may not always be present.

In a co-algebraic treatment based on an underlying category of ultrametric spaces, carried out by de Vink and Rutten [dVR99], they were able to show that weak pullbacks do exist in their category. The Giry monad was used to define LMPs as coalgebras and zigzags are precisely the coalgebra homomorphisms. Unfortunately, common spaces like the reals are not ultrametric spaces.

In the paper by Desharnais *et al.* [DEP02] an even weaker construction called the semi-pullback was used. This does allow one to complete the square but there are numerous technical details to verify [Eda99] and the construction requires one to jump back and forth between different categories. It is the complexity of this construction that has led to the search for other approaches, finally leading to the relatively elementary treatment of this book, which first appeared in [DGJP03].

A much deeper analysis of the semi-pullback construction using more powerful mathematical tools appeared shortly thereafter [Dob03]. This approach, due to Doberkat, has been developed by him in several recent papers and in a forthcoming book. The main advances are that he can work in relatively civilised categories and does not have to move between different settings. The main background needed for following his work is descriptive set theory [Mos80; Kec95; Sri98], particularly so-called selection theorems that give techniques for inverting measurable functions.

13.2 Metrics

The idea that one should work with metrics rather than equivalence relations goes back to Jou and Smolka [JS90]. The first construction of a pseudo-metric whose kernel is probabilistic bisimulation is due to Desharnais, Gupta, Jagadeesan and the present author [DGJP99; DGJP04]. The algorithm presented there is extremely inefficient and merely shows that the metric is computable.

The work of van Breugel and Worrell [vBW01] used the connection with linear programming to present a polynomial-time algorithm for computing the metric, at least in the case where the discount factor is less than 1. This turned out to be a very fruitful connection and duality ideas from linear programming were much used.

Ferns et al. [FPP04; FPP05] showed how these ideas could be adapted to Markov decision processes (MDPs) and established bounds for how far the value functions could be from optimal in terms of the metric distance. This led to the hope that the metric could be used for applications in AI planning. Unfortunately, the algorithm was still not efficient enough for applications to planning in large unstructured MDPs. Ferns *et al.* [FCPP06; Fer08] developed techniques based on approximating the metric by sampling which hold some promise for applications.

13.3 Nondeterminism

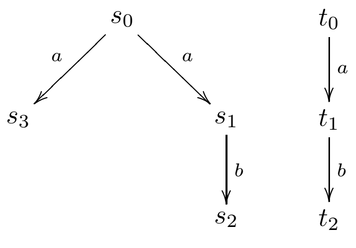
In this book the focus has been exclusively on what are called "fully probabilistic" processes. For many applications, especially in concurrency, it is necessary to consider nondeterminism. One may just not have the data for a fully probabilistic model. There are two main ways of thinking of combining probability and nondeterminism. These are called the *alternating* model and the *probabilistic automaton* model.

In the alternating model the state space is partitioned into two sets called S_p and S_n. The states in S_p can only take probabilistic transitions while the states in S_n take the nondeterministic transitions. Despite the name, it is not required that the states strictly alternate but, for convenience, it is usually assumed that they do. In essence, the alternating model allows one to *name* probability distributions. This means that when one matches with respect to bisimulation one requires that the distributions named by the states in S_p must also match.

The alternating model arises from a model proposed by Vardi in 1985 [Var85]. The explicit description of it is due to Hansson and Jonsson [HJ90; HJ94a] and it has also been used in the discussion of weak bisimulation, which we address below.

In the probabilistic automaton model, due to Segala and Lynch [SL94], there is no distinction made between two kinds of states. There are two choices involved in a transition. The choice of a label is made, perhaps by the environment, and then a scheduler chooses a transition probability function. The distribution of final states is determined by these two choices together. Thus, for each label, there is not necessarily a unique transition probability function, the choice is limited by the label chosen but there is still some freedom left for the scheduler. This is more like the labelled transition systems that arise when one is dealing with non-probabilistic process algebra. Matching now has to take place relative to the possible schedulers. In fact, the schedulers can be randomised. The notions of bisimulation are incomparable between the alternating model and the probabilistic automaton model.

Logical characterisation cannot work with just the negation free formulas that we used for LMPs when one has the mixture of probability and nondeterminism. Even in the non-probabilistic case some negative formulas are required in order to characterise bisimulation. A well-known example is the given by the following pair of processes:



The states s_0 and t_0 are not bisimilar but no negation-free formula can tell them apart.

In the probabilistic automata model there are two closely related concepts: strong bisimulation and strong probabilistic bisimulation. The key difference is that in strong bisimulation one requires that transitions match but with the strong probabilistic bisimulation one allows matching with convex combinations of transitions. In the alternating model they coincide, but in the PA model they are different. A good discussion of all this is contained in a review paper by Segala [Seg06].

The question of a good mathematical model combining probability and nondeterminism has been considered from a domain-theoretic point of view. One can define a coalgebra for nondeterminism using a powerset or powerdomain monad and another coalgebra for probabilistic transition systems using the Gìry monad. Combining these monads is troublesome however because there they do not distribute over each other and it is not clear what equations should hold. Examples of investigations along these lines are due to Mislove *et al.* [MOW03] and in unpublished work by Plotkin and Keimel.

13.4 Testing

One of the criticisms of bisimulation is that it does not capture how systems are used in various contexts. Testing equivalence captures this and was explicitly designed to be a more intuitive account of process equivalence. The paper by Larsen and Skou [LS91] does address testing equivalence and shows that there is a close relation between testing equivalence and bisimulation, indeed in a more natural way than what had emerged for ordinary nondeterministic processes [Abr91a].

The theory of testing for LMPs and a characterisation of bisimulation in terms of testing was given by van Breugel et al. [vBMOW05]. The testing processes in that case are very similar to the formulas appearing in the logical characterisation. A very appealing duality theory for LMPs was based on these ideas [MMW04].

Testing equivalences for probabilistic automata are much more subtle; a very insightful analysis was given by Stoelinga and Vaandrager [SV03].

13.5 Weak Bisimulation

In many applications one wants to view certain actions as internal to a system and hence unobservable. The notion of bisimulation equivalence is then defined as before but one is allowed to ignore unobservable actions. For fully probabilistic systems the notion of weak bisimulation was developed by Baier and Hermanns [BH97]. In the mixed probabilistic-nondeterministic case one has to be careful about defining weak bisimulation because the unobservable actions may or may not cause one to leave a bisimulation equivalence class.

As before, weak bisimulation can be studied in an alternating model [PLS00] or with probabilistic automata [SL94]. In the case of the alternating model, Philippou *et al.* used conditioning to capture the idea that the unobservable actions may or may not take one out of a bisimulation equivalence class. It turns out that one can work instead with convex combinations of actions and develop an equivalent notion of weak bisimulation [DGJP02b]. In that version, one can also prove a logical characterisation theorem for weak bisimulation but there are new subtleties. The main point is that one is no longer working with probability distributions but with sets of probability distributions and taking the supremum. This yields a quantity called a capacity which, unlike distributions, is not additive. The proofs thus had to use new techniques – essentially based on continuity properties – in order to proceed.

Just as one can define pseudometrics whose kernels give bisimulation, one can define pseudometrics whose kernels give weak bisimulation. The metric analogue of weak bisimulation was developed by Desharnais *et al.* [DGJP02a] using many of the ideas of van Breugel and Worrell on linear programming [vBW01].

13.6 Approximation

The approximation techniques in the present book all use the idea that the approximants should under-approximate the transition probabilities; this allows a good match with simulation and the logic. There is, however, a natural approach which is to average the transition probabilities. The natural tool for computing an average is the conditional expectation. A preliminary study of conditional expectation appears in a paper by Danos *et al.* [DDP03].

The main idea is that one would like to work with a cruder version of the given σ-algebra so that one does not have to consider transition probabilities to all the sets of the original σ-algebra. Let us suppose that the given σ-algebra is Σ but we wish to work with a subalgebra $\Lambda \subseteq \Sigma$. Then we can work with the conditional expectation given Λ. Intuitively, the conditional expectation averages over the cells of Λ and ignores the fine structure revealed by Σ.

The major subtlety there is that conditional expectations are only defined uniquely "up to a set of measure 0". This is awkward if one is interested in actually computing with them. There is, however, a condition which was called *granularity* in [DDP03] according to which one gets unique condition expectations. This is a very strong condition and requires that the system come with a canonical measure of some kind. For many applications this is reasonable; there is usually some close relative of the Lebesgue measure in the picture. However, the general theory is still under development.

13.7 Model Checking

There is a huge literature on model checking for probabilistic and real-time systems and it would be futile to even attempt a survey here. There is an excellent recent text book [BK08] presentation of the subject of model checking which includes a substantial discussion of model checking probabilistic systems: this is commonly – but misleadingly – called probabilistic model checking.

Probabilistic model checking began with a paper by Bianco and de Alfaro [BdA95]. There was an explosion of subsequent work, much of it associated with the names of Baier, Clark, de Alfaro, Hermanns, Huth, Katoen, Kwiatkowska, Norman, Parker, Segala and many others [Bai96; BCHG$^+$97; HK97; dAKN$^+$00; KNP04].

The subject has gone beyond the realm of theory. A successful system called PRISM has been built by Marta Kwiatkowska and her group and a number of industrial case studies have been carried out with PRISM [KNP05].

One of the interesting developments in the field is the convergence of interest between the performance evaluation community and the probabilistic verification community. The performance evaluation community has long been interested in quantitative reasoning about large systems and ideas for compact representations of large systems and compositional techniques have moved back and forth between the two communities. Of particular note is Performance Evaluation Process Algebra (PEPA) developed by Jane Hillston [Hil94] which is a compositional approach to performance evaluation based on ideas from process algebra. Techniques from the performance evaluation community have led to new model checking techniques for real-time systems and other properties of interest to performance evaluation [BHHK00; HCH$^+$02].

Bibliography

[Abr91a] S. Abramsky. A domain equation for bisimulation. *Information and Computation*, 92(2):161–218, 1991.

[Abr91b] S. Abramsky. Domain theory in logical form. *Annals of Pure and Applied Logic*, 51:1–77, 1991.

[AMESD98] M. Alvarez-Manilla, A. Edalat, and N. Saheb-Djahromi. An extension result for continuous valuations. *Electronic Notes in Theoretical Computer Science*, 13, 1998.

[AN87] E. J. Anderson and P. Nash. *Linear Programming in Infinite-dimensional Spaces.* Discrete Mathematics and Computation. Wiley-Interscience, 1987.

[Arv76] W. Arveson. *An Invitation to C^*-Algebra.* Springer-Verlag, 1976.

[Ash72] R. B. Ash. *Real Analysis and Probability.* Academic Press, 1972.

[ASSB96] Adnan Aziz, Kumud Sanwal, Vigyan Singhal, and Robert K. Brayton. Verifying continuous time Markov chains. In *Proc. of Conference on Computer-Aided Verification*, 1996.

[Bai] C. Baier. Zeno paths occur with probability zero. Private communication.

[Bai96] C. Baier. Polynomial time algorithms for testing probabilistic bisimulation and simulation. In *Proceedings of the 8th International Conference on Computer Aided Verification (CAV'96)*, number 1102 in Lecture Notes in Computer Science, pages 38–49, 1996.

[BCFPP05] Alexandre Bouchard-Côté, Norm Ferns, Prakash Panangaden, and Doina Precup. An approximation algorithm for labelled markov processes: towards realistic approximation. In *Proceedings of the 2nd International Conference on the Quantitative Evaluation of Systems (QEST)*, pages 54–61, September 2005.

[BCHG$^+$97] Christel Baier, Ed Clark, Vasiliki Hartonas-Garmhausen, Marta Kwiatkowska, and Mark Ryan. Symbolic model checking for probabilistic processes. In *Proceedings of the 24th International Colloquium On Automata Languages And Programming*, number 1256 in Lecture Notes In Computer Science, pages 430–440, 1997.

[BdA95] A. Bianco and L. de Alfaro. Model checking of probabilistic and nondeterministic systems. In P. S. Thiagarajan, editor, *Proceedings of the 15th Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, number 1026 in Lecture Notes In Computer Science, pages 499–513, 1995.

[BDEP97] R. Blute, J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation for labelled Markov processes. In *Proceedings of the Twelfth IEEE Symposium On Logic In Computer Science, Warsaw, Poland.*, 1997.

[Ber80] J. O. Berger. *Statistical Decision Theory.* Springer-Verlag, 1980.

[BH97] C. Baier and H. Hermanns. Weak bisimulation for fully probabilistic processes. In *Proceedings of the 1997 International Conference on Computer Aided Verification*, number 1254 in Lecture Notes In Computer Science. Springer-Verlag, 1997.

[BHHK00] C. Baier, B. Haverkort, H. Hermanns, and J.-P. Katoen. Model checking continuous-time Markov chains by transient analysis. In *CAV 2000: Proceedings of the 12th Annual Symposium on Computer-Aided Verification*, number 1855 in Lecture Notes In Computer Science, pages 358–372. Springer-Verlag, 2000.

[Bil95] P. Billingsley. *Probability and Measure.* Wiley-Interscience, 1995.

[BK08] Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking.* MIT Press, 2008.

[Bre68] L. Breiman. *Probability.* Addison-Wesley, 1968.

[Cho53] G. Choquet. Theory of capacities. *Ann. Inst. Fourier (Grenoble)*, 5:131–295, 1953.

[Chv83] Vasek Chvatal. *Linear Programming.* W. H. Freeman and Company, 1983.

[CM65] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes.* Chapman and Hall, 1965.

[dA97] L. de Alfaro. *Formal Verification of Probabilistic Systems.* PhD thesis, Stanford University, 1997. Technical Report STAN-CS-TR-98-1601.

[dAHM03] L. de Alfaro, T. Henzinger, and R. Majumdar. Discounting the future in systems theory. In J. Baeten, J. K. Lenstra, J. Parrow, and G. J. Woeginger, editors, *Thirtieth International Colloquium On Automata Languages And Programming*, number 2719 in Lecture Notes In Computer Science, pages 1022–1037. Springer-Verlag, 2003.

[dAKN$^+$00] L. de Alfaro, M. Kwiatkowska, G. Norman, D. Parker, and R. Segala. Symbolic model checking of concurrent probabilistic processes using MTBDDs and the Kronecker representation. In *Proceedings of Tools and Algorithms for the Construction and Analysis of Systems 2000*, number 1785 in Lecture Notes In Computer Science, pages 395–410. Springer-Verlag, 2000.

[DD03a] Vincent Danos and Josée Desharnais. A fixpoint logic for labeled Markov Processes. In Zoltan Esik and Igor Walukiewicz, editors, *Proceedings of an international Workshop FICS'03 (Fixed Points* *in Computer Science)*, Warsaw, 2003.

[DD03b] Vincent Danos and Josée Desharnais. Labeled Markov Processes: Stronger and faster approximations. In *Proceedings of the 18th Symposium on Logic in Computer Science*, Ottawa, 2003. IEEE.

[DDP03] Vincent Danos, Josée Desharnais, and Prakash Panangaden. Conditional expectation and the approximation of labelled markov processes. In Roberto Amadio and Denis Lugiez, editors, *CONCUR 2003 - Concurrency Theory*, volume 2761 of *Lecture Notes In Computer Science*, pages 477–491. Springer-Verlag, 2003.

[DDP04] Vincent Danos, Josée Desharnais, and Prakash Panangaden. Labelled markov processes: Stronger and faster approximations. *Electronic Notes in Theoretical Computer Science*, 87:157–203, November 2004.

[DEP98] J. Desharnais, A. Edalat, and P. Panangaden. A logical characterization of bisimulation for labelled Markov processes. In *proceedings of the 13th IEEE Symposium On Logic In Computer Science, Indianapolis*, pages 478–489. IEEE Press, June 1998.

[DEP02] J. Desharnais, A. Edalat, and P. Panangaden. Bisimulation for labeled Markov processes. *Information and Computation*, 179(2):163–193, Dec 2002.

[Des99] J. Desharnais. *Labelled Markov Processes.* PhD thesis, McGill University, November 1999.

[DGJP99] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Metrics for labeled Markov systems. In *Proceedings of CONCUR99*, number 1664 in Lecture Notes in Computer Science. Springer-Verlag, 1999.

[DGJP00] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Approximation of labeled Markov processes. In *Proceedings of the Fifteenth Annual IEEE Symposium On Logic In Computer Science*, pages 95–106. IEEE Computer Society Press, June 2000.

[DGJP02a] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. The metric analogue of weak bisimulation for labelled Markov processes. In *Proceedings of the Seventeenth Annual IEEE Symposium On Logic In Computer Science*, pages 413–422, July 2002.

[DGJP02b] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Weak bisimulation is sound and complete for *pctl*∗. In L. Brim, P. Jancar, M. Kretinsky, and A. Kucera, editors, *Proceedings of 13th International Conference on Concurrency Theory, CONCUR02,*, number 2421 in Lecture Notes In Computer Science, pages 355–370. Springer-Verlag, 2002.

[DGJP03] J. Desharnais, V. Gupta, R. Jagadeesan, and P. Panangaden. Approximating labeled Markov processes. *Information and Computation*, 184(1):160–200, July 2003.

[DGJP04] Josée Desharnais, Vineet Gupta, Radhakrishnan Jagadeesan, and Prakash Panangaden. A metric for labelled Markov processes. *Theoretical Computer Science*, 318(3):323–354, June 2004.

[Dil88] D. Dill. *Trace Theory for Automatic Hierarchical Verification of Speed-Independent Circuits.* ACM Distinguished Dissertations. MIT Press, 1988.

[Dob03] E.-E. Doberkat. Semi-pullbacks and bisimulations in categories of stochastic relations. In J. C. M. Baeten, J. K. Lenstra, J. Parrow, and G. J. Woeinger, editors, *Proceedings of the 27th International Colloquium On Automata Languages And Programming, ICALP'03*, number 2719 in Lecture Notes In Computer Science, pages 996–1007. Springer-Verlag, July 2003.

[DP03] Josée Desharnais and Prakash Panangaden. Continuous stochastic logic characterizes bisimulation for continuous-time Markov processes. *Journal of Logic and Algebraic Progamming*, 56:99–115, 2003. Special issue on Probabilistic Techniques for the Design and Analysis of Systems.

[Dud89] R. M. Dudley. *Real Analysis and Probability.* Wadsworth and Brookes/Cole, 1989.

[dVR99] E. de Vink and J. J. M. M. Rutten. Bisimulation for probabilistic transition systems: A coalgebraic approach. *Theoretical Computer Science*, 221(1/2):271–293, June 1999.

[Eda99] Abbas Edalat. Semi-pullbacks and bisimulation in categories of Markov processes. *Mathematical Structures in Computer Science*, 9(5):523–543, 1999.

[FCPP06] N. Ferns, P. Castro, P. Panangaden, and D. Precup. Methods for computing state similarity in Markov decision processes. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 174–181, 2006.

[Fel71] W. Feller. *An Introduction to Probability Theory and its Applications II.* John Wiley and Sons, 2nd edition, 1971.

[Fer08] Norm Ferns. *State-Similarity Metrics for Continuous Markov Decision Processes.* PhD thesis, McGill University, 2008.

[FPP04] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision precesses. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 162–169, July 2004.

[FPP05] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for markov decision processes with infinite state spaces. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 201–208, July 2005.

[Ger85] Robert Geroch. *Mathematical Physics.* Chicago Lectures in Physics. University of Chicago Press, 1985.

[Gir81] M. Giry. A categorical approach to probability theory. In B. Banaschewski, editor, *Categorical Aspects of Topology and Analysis*, number 915 in Lecture Notes In Mathematics, pages 68–85. Springer-Verlag, 1981.

[GJP04] Vineet Gupta, Radhakrishnan Jagadeesan, and Prakash Panangaden. Approximate reasoning for real-time probabilistic processes.

In *The Quantitative Evaluation of Systems, First International Conference QEST04*, pages 304–313. IEEE Press, 2004.

[GJP06] Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. Approximate reasoning for real-time probabilistic processes. *Logical Methods in Computer Science*, 2(1):paper 4, 2006.

[GKK$^+$80] G.Gierz, K.H.Hoffman, K.Keimel, J.D.Lawson, M.Mislove, and D.S.Scott, editors. *A compendium of continuous lattices*. Springer-Verlag Berlin Heidelberg New York, 1980.

[GKK$^+$03] G.Gierz, K.H.Hoffman, K.Keimel, J.D.Lawson, M.Mislove, and D.S.Scott. *Continuous lattices and domains*. Number 93 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2003.

[Hal74] P. Halmos. *Measure Theory*. Number 18 in Graduate Texts in Mathematics. Springer-Verlag, 1974. Originally published in 1950.

[HCH$^+$02] B. R. Haverkort, L. Cloth, H. Hermans, J.-P. Katoen, and C. Baier. Model checking performability properties. In *Proceedings of the International COnference on Dependable Systems and Networks 2002*, pages 102–113. IEEE Computer Society, June 2002.

[Her02] H. Hermanns. *The Quest for Quantified Quality*. Number 2428 in Lecture Notes In Computer Science. Springer-Verlag, 2002.

[Hil94] J. Hillston. *A Compositional Approach to Performance Modelling*. PhD thesis, University of Edinburgh, 1994. Published as a Distinguished Dissertation by Cambridge University Press in 1996.

[HJ90] H. Hansson and B. Jonsson. A calculus for communicating systems with time and probabilities. In *Proceedings of the 11th IEEE Real-Time Systems Symposium*, pages 278–287. IEEE Computer Society Press, 1990.

[HJ94a] H. Hansson and B. Jonsson. A logic for reasoning about time and reliability. *Formal Aspects of Computing*, 6(5):512–535, 1994.

[HJ94b] J. Hoffman-Jørgenson. *Probability With a View Towards Applications - 2 volumes*. Chapman and Hall, 1994.

[HK97] M. Huth and M. Kwiatkowska. Quantitative analysis and model checking. In *proceedings of the 12 IEEE Symposium On Logic In Computer Science*, pages 111–122. IEEE Press, 1997.

[Hut81] J. E. Hutchinson. Fractals and self-similarity. *Indiana Univ. Math. J.*, 30:713–747, 1981.

[Jec71] Thomas J. Jech. *Lectures in Set Theory*. Number 217 in Lecture Notes In Mathematics. Springer-Verlag, 1971.

[JL91] B. Jonsson and K. Larsen. Specification and refinement of probabilistic processes. In *Proceedings of the 6th Annual IEEE Symposium On Logic In Computer Science*, 1991.

[Jon90] C. Jones. *Probabilistic Non-determinism*. PhD thesis, University of Edinburgh, 1990. CST-63-90.

[Jos92] M. B. Josephs. Receptive process theory. *Acta Informatica*, 29(1):17–31, February 1992.

[JP89] C. Jones and G. D. Plotkin. A probabilistic powerdomain of evaluations. In *Proceedings of the Fourth Annual IEEE Symposium On Logic In Computer Science*, pages 186–195, 1989.

[JS90] C.-C. Jou and S. A. Smolka. Equivalences, congruences, and complete axiomatizations for probabilistic processes. In J.C.M. Baeten and J.W. Klop, editors, *CONCUR 90 First International Conference on Concurrency Theory*, number 458 in Lecture Notes In Computer Science. Springer-Verlag, 1990.

[JT98] A. Jung and R. Tix. The troublesome probabilistic powerdomain. *Electronic Notes in Theoretical Computer Science*, 13, 1998.

[Jun88] A. Jung. *Cartesian Closed Categories of Domains.* PhD thesis, Technischen Hogschule Darmstadt, 1988.

[Kec95] Alexander S. Kechris. *Classical Descriptive set Theory*, volume 156 of *Graduate Texts in Mathematics.* Springer-Verlag, 1995.

[KNP04] M. Kwiatkowska, G. Norman, and D. Parker. *Mathematical Techniques for Analyzing Concurrent and Probabilistic Systems*, chapter Modelling and Verification of Probabilistic Systems, pages 93–215. Number 23 in CRM Lecture Notes. American Mathematical Society, 2004.

[KNP05] M. Kwiatkowska, G. Norman, and D. Parker. Probabilistic model checking in practice: Case studies with prism. *ACM Performance Evaluation Review*, 32(4):16–21, March 2005.

[Koz81] D. Kozen. Semantics of probabilistic programs. *Journal of Computer and Systems Sciences*, 22:328–350, 1981.

[Koz85] D. Kozen. A probabilistic PDL. *Journal of Computer and Systems Sciences*, 30(2):162–178, 1985.

[KS60] J. G. Kemeny and J. L. Snell. *Finite Markov Chains.* Van Nostrand, 1960.

[KT66] J. F. C. Kingman and S. J. Taylor. *Introduction to Measure and Probability.* Cambridge University Press, 1966.

[Law63] F. W. Lawvere. Functorial semantics of algebraic theories. *Proc. Nat. Acad. Sci. U.S.A.*, 50:869–872, 1963.

[Law82] J. D. Lawson. Valuations on continuous lattices. In R.-E. Hoffman, editor, *Continuous lattices and related topics*, volume 27 of *Mathematik Arbeitspapiere*, pages 204–225. Universität Bremen, 1982.

[Law97] J. Lawson. Spaces of maximal points. *Mathematical Structures in Computer Science*, 7(5):543–555, October 1997.

[LS91] K. G. Larsen and A. Skou. Bisimulation through probablistic testing. *Information and Computation*, 94:1–28, 1991.

[LT89] N. A. Lynch and M. R. Tuttle. An introduction to input/output automata. *CWI Quarterly*, 2(3):219–246, 1989.

[MA86] E. Manes and M. Arbib. *Algebraic Approaches to Program Semantics.* Springer-Verlag, 1986.

[Mil80] R. Milner. *A Calculus for Communicating Systems*, volume 92 of *Lecture Notes in Computer Science.* Springer-Verlag, 1980.

[MMW04] D. Pavlovic M. Mislove, J. Ouaknine and J. Worrell. Duality for labelled markov processes. In I. Walukiewicz, editor, *Foundations of*

Software Science and Computation Structures, FOSSACS, volume 2987 of *Lecture Notes In Computer Science*, pages 393–407, 2004.

[Mos80] Yiannis N. Moschovakis. *Descriptive Set Theory.* Elsevier North Holland, 1980.

[MOW03] M. W. Mislove, J. Ouaknine, and J. B. Worrell. Axioms for probability and nondeterminism. In *Proceedings of EXPRESS 2003*, 2003. Appeared as an issue of ENTCS.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[Pac78] J. Pachl. Disintegration and compact measures. *Math. Scand.*, 43:157–168, 1978.

[Pan99] Prakash Panangaden. The category of Markov processes. *ENTCS*, 22:17 pages, 1999. `http://www.elsevier.nl/locate/entcs/volume22.html`.

[Par67] K. R. Parthasarathy. *Probability Measures on Metric Spaces.* Academic Press, 1967.

[Par81] David Park. Title unknown. Slides for Bad Honnef Workshop on Semantics of Concurrency, 1981.

[PLS00] A. Philippou, I. Lee, and O. Sokolsky. Weal bisimulation for probabilistic processes. In C. Palamidessi, editor, *Proceedings of CONCUR 2000*, number 1877 in Lecture Notes In Computer Science, pages 334–349. Springer-Verlag, 2000.

[Pop94] Sally Popkorn. *First Steps in Modal Logic.* Cambridge University Press, 1994.

[PS85] J. L Peterson and A. Silberschatz. *Operating System Concepts.* Addison-Wesley Inc., 1985.

[Put94] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, 1994.

[RdV97] J. J. M. M. Rutten and E. de Vink. Bisimulation for probabilistic transition systems: a coalgebraic approach. In P. Degano, editor, *Proceedings of ICALP 97*, number 1256 in Lecture Notes In Computer Science, pages 460–470. Springer-Verlag, 1997.

[Rud66] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, 1966.

[San04] Davide Sangiorgi. Bisimulation and coinduction: from the origins to today. slides for an Invited Talk at LICS 2004, 2004.

[SD80] N. Saheb-Djahromi. Cpos of measures for nondeterminism. *Theoretical Computer Science*, 12(1):19–37, 1980.

[Seg06] R. Segala. Probability and nondeterminism in operational models of concurrency. In *Proceedings of the 17th International Conference on Concurrency Theory CONCUR '06*, number 4137 in Lecture Notes In Computer Science, pages 64–78, 2006.

[SL94] R. Segala and N. Lynch. Probabilistic simulations for probabilistic processes. In B. Jonsson and J. Parrow, editors, *Proceedings of CONCUR94*, number 836 in Lecture Notes In Computer Science, pages 481–496. Springer-Verlag, 1994.

[Sol70] R. M. Solovay. A model of set theory in which every set of reals is lebesgue measurable. *Annals of Mathematics*, 92:1–56, 1970.

[SP82] M. B. Smyth and G. D. Plotkin. The category theoretic solution of recursive domain equations. *Siam Journal of Computing*, 11(4), 1982.

[Sri98] Sashi Mohan Srivastava. *A course on Borel sets*. Number 180 in Graduate Texts in Mathematics. Springer-Verlag, 1998.

[SV03] M.I.A. Stoelinga and F.W. Vaandrager. A testing scenario for probabilistic automata. In *Proceedings of the 30th International colloquium on automata, languages and programming (ICALP'03)*, volume 2719 of *Lecture Notes in Computer Science*, pages 407–418. Springer-Verlag, 2003.

[Vaa91] F. W. Vaandrager. On the relationship between process algebra and input/output automata. In *Proceedings, Sixth Annual IEEE Symposium on Logic in Computer Science*, pages 387–398. IEEE Computer Society Press, July 1991.

[Var85] Moshe Vardi. Automatic verification of probabilistic concurrent finite-state programs. In *26th IEEE Symposium On Foundations Of Computer Science*, pages 327–338, 1985.

[vBMOW05] Franck van Breugel, Michael Mislove, Joel Ouaknine, and James Worrell. Domain theory, testing and simulation for labelled markov processes. *Theoretical Computer Science*, 333(1-2):171–197, 2005.

[vBW01] Franck van Breugel and James Worrell. An algorithm for quantitative verification of probabilistic systems. In K. G. Larsen and M. Nielsen, editors, *Proceedings of the Twelfth International Conference on Concurrency Theory - CONCUR'01*, number 2154 in Lecture Notes In Computer Science, pages 336–350. Springer-Verlag, 2001.

[Web] Russian Space Web. http://www.russianspaceweb.com/cosmos3.html. Web site.

Index