

# 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*<sup>1</sup>.

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

---

<sup>1</sup>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  $P : S \rightarrow [0, 1]$  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  $X : S \rightarrow T$ , 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  $X : S \rightarrow \mathbf{R}$  defined on the probability space  $(X, P)$  is

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

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  $\sigma$ -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<sup>2</sup>. 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  $\infty$ . 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  $i$ th 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-

---

<sup>2</sup>Modulo some minor niggling details about representing numbers uniquely.

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

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

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  $\Omega_n$  be the space of sequences of length  $n$  of heads and tails.

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

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

**Proof.**

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

□

**Proof.** (Of the weak law) We define the function  $X_j : \Omega_n \rightarrow \mathbf{R}$  by  $X_j(s) = 1$  if the  $j$ th element of the sequence  $s$  is a head and 0 if it is a tail. We define the function  $S_n : \Omega_n \rightarrow \mathbf{R}$  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

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

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  $\frac{1}{n^2} \mathbf{E}[(\sum_{i=1}^n (X_i - \frac{1}{2}))^2]$ . To calculate the expectation value we first note that if  $i$  and  $j$  are not equal, then  $\mathbf{E}[(X_i - \frac{1}{2})(X_j - \frac{1}{2})] = 0$ . Thus when we square the sum, all the cross terms have zero expectation. Thus we are left with  $\mathbf{E}[\sum_{i=1}^n (X_i - \frac{1}{2})^2]$ ; 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  $n \rightarrow \infty$  we get the result.  $\square$

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

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

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

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

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  $B_N = \bigcup_{n=N}^{\infty} A_n$ , then  $A = \bigcap_{N=1}^{\infty} B_N$ . 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  $Pr(A) = 0$ .  $\square$

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*

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

**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:

$$\leq \frac{Pr(\bigcap_{n=N}^{\infty} C_n)}{Pr(\bigcap_{n=N}^M C_n)}.$$

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

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

We used independence in the first equality above. Now if we let  $M$  go to  $\infty$  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.