Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models

by

Daniel Ray Upper

B.S. (Rice University) 1989

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Mathematics

in the

GRADUATE DIVISION

of the

UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor Morris W. Hirsch, Chair
Doctor James P. Crutchfield
Professor Jacob Feldman
Professor David R. Brillinger

1997

Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models

Copyright © 1997
by
Daniel Ray Upper

Abstract

Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models

Hidden Markov Models (HMMs) are widely used in pattern recognition applications, most notably speech recognition. However, they have been studied primarily on a practical level, with the HMM matrices as the fundamental objects and without considering the viewpoint of an observer trying to accurately predict the future output. This dissertation is a study of the processes represented by HMMs using the concepts and techniques of stochastic automata derived from the study of dynamical systems and their complexity. The goal is to understand these processes in the language of stochastic automata. Along the way, certain ideas of stochastic automata are characterized in a measure-theoretic manner.

We begin by defining a process to be a stationary measure space of bi-infinite sequences. We define a process state to be a conditional distribution on the future of a process which corresponds to the state of knowledge held by an observer who has seen some or all of the process's history. This definition is similar in spirit to ideas used in dynamical systems, and it is a formalization of the notion of a "deterministic state" used in automata theory. We describe the process states of HMM processes. And we give a necessary condition for a process to have an HMM representation.

Following [7] and [8], we define Generalized Hidden Markov Models (GHMMs). These are structurally and operationally the same as HMMs, except that parameters which are interpreted as probabilities in defining HMMs are allowed to be negative in GHMMs. We describe necessary and sufficient conditions for two GHMMs (and thus two HMMs) to represent the same process, and we give a method for finding the smallest possible GHMM equivalent to a given one.

Going further, we give an algorithm for constructing a GHMM that represents a process from the probabilities that process assigns to words. We prove that, for every process, either the algorithm constructs a GHMM that represents the given process or that no such representation exists. This characterizes the set of process representable by GHMMs. Finally, we describe an implementation of this algorithm which constructs GHMMs from sample sequences.

Contents

1 Introduction 1
2 Processes and Process States 7
2.1 Sequence Space 7
2.2 Processes 9
2.3 Past and Future 11
2.4 Histories and States 13
2.5 Process States 15
2.6 Transient and Recurrent States 20
2.7 Synchronization 23
3 Hidden Markov Models 25
3.1 Notation for Markov Chains 25
3.2 Hidden Markov Models 27
3.3 HMMs as Processes 30
3.4 Mixed States 35
3.5 Examples 43
3.6 When Does a Process have an HMM Presentation? 49
4 Generalized Hidden Markov Models 52
4.1 Generalized Hidden Markov Models 52
4.2 Redundancy and Linear Algebra 61
4.3 Equivalence and Minimization of GHMMs 72
5 Reconstruction 83
5.1 Constructing a Presentation for a Process 84
5.2 Reconstruction from a Sample 94
6 Conclusions and Further Directions 104
6.1 Process states and presentations 104
6.2 Generalized Hidden Markov Models 106
6.3 Converting GHMMs to HMMs 107
6.4 Reconstruction 110
6.5 Last remarks 110
Appendix A Notation 112
Appendix B Selected Probability Theory 116
B.1 Kolmogorov's Extension Theorem and Process Existence 116
B.2 Martingales 123
Bibliography 126

Acknowledgments

I thank my coworkers James Crutchfield, Karl Young, Jim Hanson, Lisa Borland, and Deirdre Des Jardins for the many discussions and suggestions that helped me develop this work.

I thank James Akao, William Grosso, and the members of the games group for their camaraderie and collegiality.

I thank Dorothy Brown for helping me focus on my writing.

I thank Professor Feldman and Professor Brillinger for their feedback on the various drafts of my dissertation.

I thank Professor Hirsch for giving me the freedom to take off in my own directions and work closely with Doctor Crutchfield. I would like to thank Doctor Crutchfield for his insights, advice, and enthusiasm and for understanding what I was doing. I am indebted to both of these advisors for the support, financial and otherwise, they have given me over the years.

Finally, I thank my wife, Nancy Jamieson, for her everyday presence, her emotional support, and her help in many practical ways, which together enabled me to complete this dissertation.

This work was supported at UC Berkeley by grants ONR N00014-95-1-0524 and AFOSR 91-0293 and at the Santa Fe Institute by grant ONR N00014-95-1-0975.

1 Introduction

This dissertation is about a class of stochastic processes that are usually presented as having a finite set of states, but which, in another sense, may have an infinite number of states. These processes are known variously as *Hidden Markov Models* (HMMs), functions of a Markov Chain, or stochastic finite automata, all of which are essentially equivalent. HMMs are used most widely and will be used here.

A process in the HMM class can be described as a finite-state Markov Chain with a memoryless output process which produces symbols in a finite alphabet. This is the sense in which these processes have finitely many states. However, from the perspective of an observer who knows the parameters of some representation of the process and is able to observe the output symbols but not the internal states, things look different. For some processes there are infinitely many distinct states of such an observer's knowledge about the status of the process. This knowledge is defined in terms of conditional distributions on future symbols. This is the sense in which there can be infinitely many states. These states are more relevant than the original finite set of states to the study of the process, since they allow for optimal prediction.

Functions of Markov Chains were the first descriptions of these processes to be studied, and they were initially studied as mathematical information sources [1]. There were a handful of papers such as [2] published in the 1950s and early 1960s, which define HMMs and lay out these theoretical questions. What is the entropy rate for a function of a Markov Chain? Do these two functions of Markov Chains define the same process (the identifiability question)? What is the smallest function of a Markov Chain equivalent to the given one (the minimality question)? This work was done by researchers with mathematical backgrounds, studying HMMs from a perspective of probability and information theory.

From the 1970s onward, HMMs have been used for modeling observed patterns, especially in speech recognition. There are a large number of papers, such as [3–5], that present HMMs as tools for use on these practical problems. These papers are written by researchers interested in pattern recognition, often from a viewpoint in engineering or computer science, and they usually focus on algorithms and on results in practical situations.

A new insight in 1987 led to a generalization of HMMs and to resolution of the identifiability and minimality questions about Hidden Markov Models [6]. These results appear in a very few papers that treat HMM matrices as objects of linear algebra without regard to the signs of the transition matrix entries — which have traditionally been interpreted as probabilities. These papers include [7] and [8], which solve the identifiability and minimality questions for HMMs by solving them for a generalization of HMMs known as *Generalized Hidden Markov Models* (GHMMs).

In addition to this development of HMMs per se, there is a substantial literature in computer science dealing with finite state machines. An emerging branch of this literature deals with stochastic finite automata, and falls under the heading of complex systems. This body of work, composed of papers such as [9], focuses on intrinsic processes rather than representations, and looks at stochastic automata as the simplest systems in which to study how natural systems process information. Because the definition of a stochastic finite automaton is quite similar to the definition of a Hidden Markov Model, this work provides an alternative way of looking at HMMs.

This dissertation developed from looking at HMMs from the viewpoint of the work on stochastic finite automata. This approach led to the material of chapters 2 and 3. The work in chapters 4 and 5 followed from this and used the generalization of HMMs mentioned above. The next paragraphs contain brief overviews of these chapters and are intended to give the reader an idea of what is to come.

The primary objects of chapter 2 are *processes* and *process states*. A process is a stationary probability measure on the space of bi-infinite sequences of symbols, where a *symbol* is an element of a finite set called the alphabet.* Indices into these sequences are thought of as times, as if the process were a laboratory apparatus that emits a symbol with every tick of a clock. Thus, negative indices refer to symbols that were emitted in the past and that may be known, and nonnegative indices refer to symbols that have not yet been emitted and may not have been internally determined yet. We define a word to be a finite string of symbols in the alphabet. A process assigns a probability to each word and is uniquely determined by these probabilities.

*In the stochastic process literature, the term "state" is usually used where we use "symbol". In papers such as [2,10], this leads to the somewhat confusing use of "state" to refer to both symbols and presentation states. In this dissertation, the term "state" is used exclusively to refer to objects which render the future conditionally independent of the past.

A process state is a conditional distribution over the future — a measure space of semi-infinite sequences of symbols — induced by conditioning on known historical information, such as what particular symbols were emitted at the last few ticks of the clock. The process states are the possible states of knowledge of an observer who wishes to predict the future symbols with high accuracy. This observer knows the design of our hypothetical apparatus, but not the current status of its internal components.

The definition of a process state is new in this dissertation, but the underlying idea is not. It is used, for example, in [11]. What the author has done here is to formalize this fundamental idea. The definition of a process is, of course, standard.

In chapter 3, we will introduce Hidden Markov Models (HMMs). An HMM consists of a recurrent finite-state Markov Chain, an alphabet of output symbols, and a distribution over that alphabet for each transition in the Markov Chain. The states and transitions of the Markov Chain are hidden from observation so that only the output symbols are visible. We represent an HMM primarily by a set of matrices $\left\{T^k\right\}$, one matrix T^k for each symbol k in the alphabet — note that the superscript k is an index, not an exponent. Each entry T^k_{ij} is the probability, if the Markov Chain is in state i, of emitting symbol k and going to state j. An HMM defines a process in a natural way, and so HMMs provide a convenient way to represent some processes. The states of an HMM's underlying Markov Chain are quite different from process states. We will refer to the Markov Chain's states as *presentation states*.

We compute the process states for a process represented by an HMM in terms of the presentation. We show that they are represented by probability distributions over the presentation states or, equivalently, by mixtures of presentation states. We call them *mixed states*† and we identify the real role of presentation states. The presentation states are the things we combine to make mixed states. Essentially, presentation states are basis vectors for a vector space containing the mixed states, and mixed states represent the states of knowledge of an observer. In this interpretation, the matrices T^k define linear transformations among these vectors.

Finally, we consider the question of when a process can be represented by an HMM. We prove that the following condition is necessary: if a process has an HMM presentation, then the span of the process states — a subspace of the space of conditional

†The term *mixed states* is used in the context of HMMs, with an entirely unrelated meaning in [12].

distributions on the future — is finite-dimensional. This result is a natural consequence of the fact that the process states of such a process can be represented by linear combinations of presentation states.

The definition of an HMM given in chapter 3 is one of several equivalent definitions. Mixed states first appeared in [1], although not with that name. Most of the remaining material in the chapter has not previously appeared in print, but has undoubtedly been deduced a number of times before. Stating these results in terms of process states is the author's work, as is the question, "What are the process states for a process represented by an HMM?" and the resulting interpretation of presentation states as basis vectors, both of which are crucial to this dissertation. The final section of this chapter is entirely the author's own.

Chapter 4 introduces a new class of presentations, Generalized Hidden Markov Models. We represent a GHMM like an HMM, by a matrix for each symbol, but in a GHMM we do not interpret the entries in these matrices as probabilities. Indeed, we allow them to be negative or greater than one. We do constrain these entries, and we constrain them in such a way that the calculations we use with HMMs produce meaningful results for GHMMs. In this way, a GHMM assigns probabilities to words and so it defines a process.

We justify this change as follows. For a process represented by an HMM, we calculate the probabilities of words by simple linear algebra and we represent process states as vectors in a space generated by the presentation states. But because we restrict the entries in the matrices T^k to be positive, we restrict the class of linear transformations they can represent and we restrict the set of possible mixed states to a small subset of the vector space. If we remove this constraint, we can make use of all sensible linear transformations and we can use any portion of the vector space. This is what we allow when we use GHMMs.

Two long-standing problems for HMMs can be readily solved in terms of GHMMs, namely the equivalence question — Do these two HMMs (or GHMMs) represent the same process? — and the minimization question — How small is the smallest HMM which is equivalent to a given HMM? These questions were resolved in [7]. The present work presents a new and relatively clean resolution of these questions that is similar in

spirit to the work just cited. But it is different in details and includes an important new construction, the standard presentation.

Chapter 5 contains two significant results. The first is a proof that every process such that the span of its process states is finite-dimensional can be represented as a GHMM. We show this by constructing such a representation. This completes a characterization of the processes representable by GHMMs, the first half of which appears for HMMs in chapter 3 and is generalized to GHMMs in chapter 4. The complete characterization is this: a process has a GHMM presentation if and only if the span of its process states is finite-dimensional.

It is noteworthy that we can actually construct a GHMM presentation for any process which has one. This construction leads to the second result of chapter 5: a technique for constructing a GHMM from a sample of the output from a process. This technique, which we call the *reconstruction algorithm*, is completely unlike the forward-backward algorithm, the most widely used technique for constructing HMMs from sample output. It needs further development, but it has much more solid theoretical footing than the forward-backward algorithm. Indeed, it may eventually replace the forward-backward algorithm. The work in this chapter is entirely that of the author.

In the chapter overviews, we introduced a number of concepts that may be unfamiliar to the reader. We conclude the introduction with an example, to make some of them clearer and more concrete to the reader. Let us consider the Hidden Markov Model, depicted in figure 1.1.

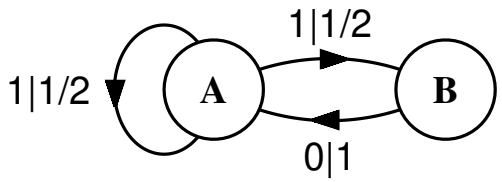

Fig. 1.1 First example HMM.

The circles **A** and **B** represent presentation states, the states of the underlying Markov Chain. The arrows connecting them represent the transitions of the HMM and the labels indicate the symbol that will be emitted if that transition is taken and the probability of that transition. For example, the label 1|1/2 at the top indicates that the symbol 1

is emitted when this transition is made and that this transition is taken on $1/2$ of the occasions when a transition is made from state **A**. We represent this same HMM with the transition matrices

$$T^0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \text{ and } T^1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 0 \end{pmatrix}. \tag{1.1}$$

This is not the full description of an HMM that we will give in chapter 3. In particular, we have not defined an initial state distribution for this HMM.

It should be clear to the reader that each presentation state in our example defines a distribution on the future symbol sequences. We call this the conditional distribution on the future given that presentation state.

Now, suppose that we know the transition matrices and can see the output of this machine, but do not know what presentation state the HMM is in. That is, the presentation states are hidden from us, whence the term *Hidden* Markov Models. If the most recent symbol we have seen is a 0, then we can deduce that the HMM is in presentation state **A**. But if the most recent symbol we have seen is a 1, we cannot deduce which presentation state the HMM is in. We can, however, infer a distribution over the presentation states. The HMM is in either state **A** or state **B** with probability $1/2$ each. And we can deduce a conditional distribution on the future given that the most recent symbol was a 1. If we do this we will find that it is equal to $1/2$ of the sum of the two conditional distributions on the future given presentation states **A** and **B**, respectively.

In the terminology of this dissertation, we have now seen two process states. The first is the conditional distribution on the future given that the most recent symbol was a 0. This happens to be identical to the conditional distribution on the future given presentation state **A**. The second is the conditional distribution on the future given that the most recent symbol was a 1, which does not correspond to either presentation state. We can represent the first of these process states by the mixed state $(P_{\mathbf{A}} = 1, P_{\mathbf{B}} = 0)$, a distribution over the presentation states which puts all probability on state **A**. Likewise, we represent the second of our process states by the mixed state $(P_{\mathbf{A}} = 1/2, P_{\mathbf{B}} = 1/2)$, a distribution which makes each presentation state equally likely.

We begin in chapter 2 with a full definition of processes and process states.

2 Processes and Process States

In this chapter and the next one, we define *processes*, the central objects of this work. We will work with them in two ways, from two different viewpoints. First we will define processes in the abstract, as probability measures on a sequence space. Second, more concretely, we will define *Hidden Markov Models* — partially observed finite state Markov chains — which we will use to represent processes. In this chapter we will introduce the first of these viewpoints, and in chapter 3 we will introduce the second and then bring the two together.

This chapter builds on, and uses the concepts and terminology of, probability theory. Appendix B contains a few necessary definitions and theorems with which the reader may be unfamiliar. A presentation of the basics, if needed, may be found in a number of standard texts, e.g. [13,14].

The reader may find it helpful to keep in mind the following metaphor. A process may be thought of as a black box on a laboratory bench with a row of lights on it. These lights are our symbols, and the row of them is our alphabet. There is a flash from one or another of these lights every second, which we describe as the emission of a symbol. This box has been running forever, and will continue running forever. We may know the design of the box, and we may have observed a number of recent symbol emissions, but we cannot observe the current configuration of the box's internal parts.

2.1 Sequence Space

In this section, we define and introduce notation for the sets upon which we will build our probability spaces.

Begin with a finite set $\mathcal{X}$ of *symbols*, which we will call an *alphabet*. Our canonical choice will be $\mathcal{X} = \{0, 1, \ldots, m-1\}$ for some natural number m. Let $\mathcal{X}^{\mathbb{Z}}$ be the space of bi-infinite sequences of elements of $\mathcal{X}$. That is, if $\mathbf{x} \in \mathcal{X}^{\mathbb{Z}}$, then $\mathbf{x}$ is a bi-infinite sequence $\ldots x_{-3}x_{-2}x_{-1}x_0x_1x_2x_3\ldots$, so that for any integer t, there is a symbol $x_t \in \mathcal{X}$. We will think of the indices as denoting measurement times, with negative indices referring to the past and nonnegative indices indicating the future. Often, we will assume that the symbols with negative indices are known and the symbols with

nonnegative indices are unknown. We will want to think of each measurement time as being either in the past or in the future, so that we will only need to consider our state of knowledge after one event an before another, and we will not need to consider our state of knowledge "as a symbol is becoming known". Thus we will be thinking of time $t = 0$ as being in the future. The reader may wish to think of the "present" as a small negative number, perhaps $-\frac{1}{2}$. Because time $t = 0$ is the smallest future time at which a symbol is observed, we will refer to x_0 as the *next symbol*.

In these terms, a *word* w of length l is an l-tuple of elements of $\mathcal{X}$, $w \in \mathcal{X}^l$. We will denote the empty word, of length zero, by λ, and the length of the word w by $|w|$. For, example, if $0, 1 \in \mathcal{X}$, $|01101| = 5$. A *subsequence* s is a structure $s = (w, (a, b))$, where w is a word and (a, b) is a pair of times such that w has length $b - a + 1$. The subsequence s is said to be an *instance* of the word w from which it is formed, and w is said to be the *base word* of the subsequence $s = (w, (a, b))$. s may also be denoted $s_a s_{a+1} \ldots s_b$. a and b are called the *start time* and *end time*, respectively, of s. The length of a subsequence is the length of its base word. An instance of λ, then, is written $(\lambda, (a, a - 1))$. A sequence $\mathbf{x}$ is said to contain, or *match*, a subsequence $s = (w, (a, b))$ if, for all $t \in \{a, \ldots, b\}$, $s_t = x_t$. We will most often refer to the *next word*, which is any subsequence $s = (w, (a, b))$ with start time $a = 0$, or the *history suffix*, which is any subsequence with end time negative one $b = -1$.

When writing a subsequence which contains x_{-1} or x_0, we will sometimes use the decimal point to denote this and to imply the start and end times. For example, $1011.$ is denotes history suffix $(1011, (-4, -1))$ and $.0110$ is denotes next word $(0110, (0, 3))$. We will not always be precise about distinguishing words from subsequences, nor about using w for words and s for subsequences. If w and z are words, then wz denotes their concatenation. The set of all words will be denoted by $\mathcal{X}^*$; this set contains λ.

The set A_s of sequences which match a subsequence s,

$$\left\{ \mathbf{x} \in \mathcal{X}^{\mathbb{Z}} | x_i = s_i \text{ for all } i \in \{a, \ldots b\} \right\} \tag{2.1}$$

is called the *cylinder set* defined by s. If s is an instance of λ, The cylinder set defined by any instance of λ is $\mathcal{X}^{\mathbb{Z}}$.

2.2 Processes

Processes are the central objects of this work. The definition of a process is this: a process is a stationary probability measure on a space of sequences. In this section we will develop this definition.

A probability measure is a function which assigns probabilities to sets — in this case, sets of sequences. To what subsets of $\mathcal{X}^{\mathbb{Z}}$ will our process assign probabilities? That is, what is the domain of this function? We need to define this set of subsets of $\mathcal{X}^{\mathbb{Z}}$, which is called a σ-field. Our choice $\mathbb{X}$ is defined to be the σ-field generated by the cylinder sets. That is, $\mathbb{X}$ is the smallest collection of subsets of $\mathcal{X}^{\mathbb{Z}}$ such that:

1. for every subsequence s, $A_s \in \mathbb{X}$, and
2. $\mathbb{X}$ is closed under complements and countable unions.

The pair $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}\right)$ is the measurable space in which we will be working.

Essentially, a probability measure on $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}\right)$ is something which assigns probabilities to the cylinder sets defined by subsequences. If s is a sequence and A_s is the cylinder set of sequences which match s, we define $\mathbf{P}(s)$ — the probability of s — to be the probability of the cylinder set $\mathbf{P}(A_s)$. Because $\mathbf{P}$ is a probability measure, $\mathbf{P}\left(\mathcal{X}^{\mathbb{Z}}\right) = 1$. Recall that $\mathcal{X}^{\mathbb{Z}} = A_\lambda$, so we have $\mathbf{P}(\lambda) = 1$.

As we stated above, we will require our processes to be stationary. A *stationary* process is one in which the probability of a subsequence does not depend on its start time. Stationarity will allow us to disregard the time index when we do not need it explicitly. In particular, it allows us to define the probability of a word to be the probability of any instance of it, as we will see shortly. Virtually all the probability measures on sequence space addressed in this work are stationary, and the exceptions will be identified as such.

The *shift map* T on a sequence space $\mathcal{X}^{\mathbb{Z}}$ is a map $T : \mathcal{X}^{\mathbb{Z}} \to \mathcal{X}^{\mathbb{Z}}$ such that for all $\mathrm{x} \in \mathcal{X}^{\mathbb{Z}}$, for all t, $(T(\mathrm{x}))_t = \mathrm{x}_{t+1}$. The shift map "shifts" a sequence over by one; it moves the time origin.

Definition 2.1.1. A process $\mathcal{P}$ is *stationary* if for all sets $A \in \mathbb{X}$, $\mathbf{P}(T(A)) = \mathbf{P}(A)$.

Note that we do not need to require shift invariance in both directions, as it is automatic. Since T is invertible in a space of bi-infinite sequences, let $B = T^{-1}(A)$, and apply the definition of stationarity to B, and we have $\mathbf{P}\left(T^{-1}(A)\right) = \mathbf{P}(A)$.

Definition 2.1.2. A *process* $\mathcal{P}$ is a stationary probability space $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$. That is, $\mathcal{P}$ is a stationary probability measure $\mathbf{P}$ on the measurable space $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}\right)$.

The author has chosen to work exclusively with stationary processes for convenience and for reasons which have their roots in the historical development of this work. However, virtually all of the results in this dissertation are valid with a definition which does not require processes to be stationary and instead expects a process to have a start and stop at finite times. Some of the statements made here, and many of the proofs, require modifications in order to apply under such a definition.

Let w be a word and let s be an instance of w. If $\mathbf{P}$ is stationary, all instances of w have the same probability. Thus, we can define the probability of w to be $\mathbf{P}(w) = \mathbf{P}(s)$, so we can think of a process $\mathcal{P}$ as a function which assigns probabilities to words. We will use W_l to denote the set of all words of length l with positive measure. This leads us to two facts, which are trivial consequences of the properties of probability measures. The first of these is that, because the cylinder set induced by λ is defined to be all of $\mathcal{X}^{\mathbb{Z}}$,

$$\mathbf{P}(\lambda) = 1. \tag{2.2}$$

The second is that, for any word w and any length $l \geq 0$,

$$\sum_{z \in W_l} \mathbf{P}(wz) = \sum_{z \in W_l} \mathbf{P}(zw) = \mathbf{P}(w). \tag{2.3}$$

As it happens, the converse of this pair of facts is true — any function on $\mathcal{X}^*$ which satisfies equations 2.2 and 2.3 defines a process. This will be our primary tool for showing that a particular object is a stationary process.

In the statement of the theorem B.1.1, we have separated f from $\mathbf{P}$ for clarity. When we use this theorem , we will not usually mention f explicitly. Instead, we will use $\mathbf{P}$ in the role which f serves here, verify equations 2.2 and 2.3, and invoke the theorem to assert that $\mathbf{P}$ describes a unique process. This renders the distinction between f and $\mathbf{P}$ moot.

Theorem B.1.1. Given a map $f : \mathcal{X}^* \to [0, 1]$ satisfying

1. $f(\lambda) = 1$, and
2. For all words $w \in \mathcal{X}^*$, $f(w) = \sum_{z \in \mathcal{X}} f(zw) = \sum_{z \in \mathcal{X}} f(wz)$,

there is a unique process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ such that for all $w \in \mathcal{X}^*$, $\mathbf{P}(w) = f(w)$.

This theorem follows directly from Kolmogorov's extension theorem. Proving it is conceptually simple but requires some rather serpentine logistics. Instead of appearing here as a interruption of the conceptual development, the proof appears in appendix B.1.

A few examples are in order before we go on. We will call the first of these the *fair coin*. This process should be thought of as a bi-infinite sequence of coin flips being revealed (or flipped) as time passes. Here $\mathcal{X} = \{\text{heads}, \text{tails}\}$, and x_t is the result of the coin flip at time t. Recall that we think of negative times as the past, and nonnegative times as the future, so that we have seen all the results x_t for $t < 0$ and we have not yet seen any x_t for $t \geq 0$. For any word w of length l, $\mathbf{P}(w) = 2^{-l}$. The symbols are independent and identically distributed (iid) so the process must be stationary. Every sequence in $\mathcal{X}^{\mathbb{Z}}$ is a realization (defined in section 2.3) of the fair coin process. Although we don't need theorem B.1.1 to prove that this process exists, we will verify that equations 2.2 and 2.3 are satisfied. For 2.2, $f(\lambda) = 2^0 = 1$. For 2.3, if w has length l and $z \in \mathcal{X}$, then $f(wz) = f(zw) = 2^{-(l+1)}$, so we have $2 \cdot 2^{-(l+1)} = 2^l = f(w)$.

The second of our examples is a strictly periodic process, in which a fixed word is repeated over and over. In this example, we choose the word 10000, and a typical realization looks like

$$\ldots 0001000010000100001000010000100\ldots \tag{2.4}$$

If the phase (i.e. the index associated to a 1, taken modulo 5) is uniformly distributed, $\mathbf{P}$ is stationary, and we have a process. It is not difficult to see that every word is assigned a probability of either 0 or $\frac{1}{5}$, with the exception of λ, 0, 00, and 000, which have probabilities 1, $\frac{4}{5}$, $\frac{3}{5}$, and $\frac{2}{5}$, respectively. This is clearly a process — the measure can be described explicitly. It assigns measure $\frac{1}{5}$ each to five bi-infinite sequence and 0 to all the others.

2.3 Past and Future

At times, we will need to treat the past and the future separately. In this section we will introduce a decomposition of a processes underlying probability space $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}\right)$ which will facilitate this.

Let $\mathcal{X}^+$, the *future space*, be the usual set of semi-infinite sequences of elements of $\mathcal{X}$,

$$\mathcal{X}^+ = \left\{ \mathbf{x}^+ = x_0, x_1, \ldots \,|\, \text{for all } i \geq 0,\ x_i \in \mathcal{X} \right\} \tag{2.5}$$

and $\mathcal{X}^-$, the *history space*, be the set of semi-infinite sequences of elements of $\mathcal{X}$ with negative indices:

$$\mathcal{X}^- = \left\{ \mathbf{x}^- = \ldots, x_{-2}, x_{-1} |\, \text{for all } i < 0,\ x_i \in \mathcal{X} \right\} \tag{2.6}$$

We will think of a bi-infinite sequence $\mathbf{x} \in \mathcal{X}^{\mathbb{Z}}$ as an ordered pair of semi-infinite sequences $\left(\mathbf{x}^-, \mathbf{x}^+\right)$, where $\mathbf{x}^- \in \mathcal{X}^-$ and $\mathbf{x}^+ \in \mathcal{X}^+$. Thus, we can write $\mathcal{X}^{\mathbb{Z}}$ as a product of sets, $\mathcal{X}^{\mathbb{Z}} = \mathcal{X}^- \times \mathcal{X}^+$. We will refer to an element $\mathbf{x}^-$ of $\mathcal{X}^-$ as a *history*, and an element $\mathbf{x}^+$ of $\mathcal{X}^+$ as a *future*. Thus a history suffix, which we defined in section 2.1 to be a subsequence with end time -1, is in fact a suffix of a history.

Next, we will define the *history σ-field* $\mathbb{H}$ and the *future σ-field* $\mathbb{F}$, both of which are σ-fields on $\mathcal{X}^{\mathbb{Z}}$ and are sub-σ-fields of $\mathbb{X}$. $\mathbb{H}$ and $\mathbb{F}$ are generated by the cylinder sets in $\mathcal{X}^-$ and $\mathcal{X}^+$ respectively. That is, $\mathbb{H}$ is defined to be the σ-field on $\mathcal{X}^{\mathbb{Z}}$ generated by all cylinder sets defined by subsequences with negative end times, and $\mathbb{F}$ is defined to be the σ-field on $\mathcal{X}^{\mathbb{Z}}$ generated by all cylinder sets with nonnegative start times. Thus if $s = (w, (a, b))$ is a subsequence with end time $b \leq -1$, then A_s — the set of all bi-infinite sequences $\mathbf{x} \in \mathcal{X}^{\mathbb{Z}}$ which match s — is an event in $\mathbb{H}$. We will refer to elements of $\mathbb{H}$ as *history events*. Intuitively, a history event is a set for which membership depends only on the history part of a sequence. A similar statement is true for $\mathbb{F}$, and we will refer to elements of $\mathbb{F}$ as *future events*.

In addition, we will need to define a series of finite history σ-fields $\{\mathbb{H}_l | l \in \{0, 1, \ldots\}\}$, where $\mathbb{H}_l$ is the σ-field on $\mathcal{X}^{\mathbb{Z}}$ generated by all length l history suffixes. ($\mathbb{H}_0$ is the trivial σ-field on $\mathcal{X}^{\mathbb{Z}}$.) An event in $\mathbb{H}_l$, then, is the set of bi-infinite sequences $\mathbf{x}$ which satisfy some in positions $x_{-l}, \ldots, x_{-1}$. As the reader may readily verify, the finite history σ-fields form an increasing sequence. That is, for all $l \geq 0$, $\mathbb{H}_l \subset \mathbb{H}_{l+1}$. In addition, the union of the finite history σ-fields is the (infinite) history σ-field:

$$\bigcup_{l=0}^{\infty} \mathbb{H}_l = \mathbb{H}. \tag{2.7}$$

A *realization* from a process $\mathcal{P}$ is a bi-infinite sequence in $\mathcal{X}^{\mathbb{Z}}$ which, loosely speaking, is within the support of the process' measure $\mathbf{P}$. Formally, the support of a measure depends on the topology of the measure space, and we have not defined one. Instead of doing so now, we will define a realization directly. Take a sequence $\mathrm{x} \in \mathcal{X}^{\mathbb{Z}}$, and consider the set of all subsequences which x contains. If, for each of these subsequences, the set of sequences which contains it has positive measure, then x is a realization of $\mathcal{P}$. In the same vein, a *null word* is a word with probability zero; and a history x^{-} is a *null history* if it has a suffix that is a null word.

2.4 Histories and States

In the last section, we established our notation and technical framework. The material in this section has these components, but it also has a conceptual component. We will defer most of the technical details until the next section. The material in this section draws on the concepts and terminology of Markov chains. Readers unfamiliar with Markov chains may wish to peruse section 3.1 at this time.

Consider the following situation from a time series perspective. Suppose we are watching a sequence of symbols appear as the output of an apparatus. Suppose further that this apparatus is known to be described by the process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$. Given $\mathbf{P}$ and some natural additional information — namely, the most recent few symbols — what can we predict about the future? In particular, how can we simulate future output from this apparatus?

We may assume, due to stationarity, that the next symbol will appear at time $t = 0$. Then the symbols we know form a subsequence w with stop time $b = -1$ — that is, w is a history suffix. We will make the reasonable assumption that $\mathbf{P}(w) > 0$. Then w induces a conditional distribution on x_0, $\mathbf{P}(x_0 = k|w)$.

Suppose we choose a symbol x_0 according to this conditional distribution, and build a new history suffix z from it and w shifted by one: $z_{-1} = x_0$, and $z_{t-1} = w_t$ if t is between w's start and end times. We now find ourselves in the same situation we started in, only one time step further along and with a new history suffix. We can repeat this process as many times as we like and thus (theoretically, at least) generate synthetic data.

In effect, we are using the history suffixes as states of an infinite Markov chain. If we output each new symbol as we choose it, we can accurately simulate the original

apparatus. Essentially, we've built what's called a Hidden Markov Model representation of our process. (A Hidden Markov Model is a partially observed Markov chain — see chapter 3.) However, because there are infinitely many history suffixes, it has infinitely many states, and none of these states are recurrent. That is, once a state has been seen it can never be revisited. This recasting of the process is not useful in the sense that it doesn't tell us anything new beyond knowing $\mathcal{P}$, and we have to use all the history we know in each simulation step. Is there a more concise or informative way of simulating $\mathcal{P}$?

Let's back up, and look at what we can say about distributions on the future space. Our known history suffix w induces a conditional distribution $\mathbf{P}(\cdot|w) \equiv \mathbf{P}(\cdot|A_w)$ on the future space $(\mathcal{X}, \mathbb{F})$.

$\mathbf{P}(\cdot|w)$ is an example of a *conditional future distribution*, a distribution on the future which arises when we condition on a history event. Conditional future distributions are of substantial importance. The conditional future distribution reflects our *state of knowledge* about possible future observations from $\mathcal{P}$. It includes all the information we have about the future — if we know the conditional future distribution induced by a history suffix, we may as well forget the history suffix itself. This, loosely speaking, is the sense in which $\mathbf{P}(\cdot|w)$ is a state.

In the next section, we will consider conditional future distributions which arise when we condition on either a history s or a history suffix $\mathtt{x}^{-}$. Here, however, we will restrict ourselves to distributions which arise when we condition on history suffixes. If two history suffixes, y and z, lead to the same conditional future distribution, $\mathbf{P}(\cdot|z) = \mathbf{P}(\cdot|y)$, then we will say that they are equivalent, denoted $y \sim z$. They provide the same information about the future. We define the *equivalence class* C_z of a history suffix z to be the set of all history suffixes y such that $y \sim z$: $C_z = \{y \in \mathcal{X}^* | y \sim z\}$

These equivalence classes should be thought of as condensed versions of the past, in the sense that if we remember only which equivalence class C_z a history suffix z belongs to and we forget z itself, then we have lost no information about the future. More formally, suppose we define a random variable S which maps a history suffix to its equivalence class, $S(z) = C_z$. For all future words s, the definition of C_z tells us that the conditional probability $\mathbf{P}(s|z)$ is equal to the conditional probability $\mathbf{P}(s|C_z)$. Then the history suffix and the future are conditionally independent given S.

We can now use these equivalence classes to address the problem of simulating future output. Start with equivalence class C_w which contains our known history suffix w. C_w induces a distribution on x_0. As before, we choose a symbol $k \in \mathcal{X}$ according to this distribution. Next, we choose any history suffix y in C_w — we need not choose w — and append k, and shift it over by one to get $z = yk$, which is again a history suffix. From z, we get back to an equivalence class $C_z = S(z)$. Note that any choice of $y \in C_w$ yields the same C_z, because equality of distributions over the entire future space implies equality of distributions over the subspace of sequences which start with a given x_0.

As before, we have built a Hidden Markov Model presentation of our process. This time, however, we may have gained by doing so. It is possible for the states — that is, equivalence classes — to be recurrent. For example, in the fair coin process, all history suffixes are equivalent, so there is exactly one equivalence class, and it is visited after every time step. We may have infinitely many states, or we may have only finitely many, depending on the structure of the set of equivalence classes. Essentially, if we remember only the current equivalence class, we are keeping only the information from the past that is relevant to predicting the future. This recasting, as we will see, is fundamental. These conditional distributions on the future are the basis of the primary notion of "state" that we will be using. However, because there is more than one kind of object we will want to call a state, we will refer to the equivalence classes — or rather, the conditional future distributions they induce — as *process states*.

We can extend this idea to include conditional future distributions induced by conditioning on an entire history. When $\mathbf{x}^-$ is a history and $\mathbf{P}(\cdot|\mathbf{x}^-)$ is the conditional future distribution it induces, we can compare $\mathbf{P}(\cdot|\mathbf{x}^-)$ to a conditional future distribution $\mathbf{P}(\cdot|z)$ induced by a history suffix z. Considering conditional future distributions induced by histories introduces some complications, with which we will deal in the next section.

2.5 Process States

In this section, we will go through the preceding development again, rigorously. Those readers for whom the preceding discussion constitutes an adequate definition are advised to skim the this section rather than reading it closely.

First, we will develop a rigorous definition of $\mathbf{P}(\cdot|w)$, where w is a history suffix. We are only interested in conditioning on non-null history suffixes. Let $s_l = x_{-l} \ldots x_{-1}$

and let R_l be the set of all non-null length l history suffixes, $R_l = \{s_l | \mathbf{P}(s_l) > 0\}$. Then $R = \bigcup_{l=0}^{\infty} R_l$ is the set of all non-null history suffixes. Let w be any history suffix not in R, and let $B_w \in \mathbb{H}$ be the history event containing all histories which match w. For any future event $A \in \mathbb{F}$, we define the conditional probability $\mathbf{P}(A|w)$ by Bayes rule. That is,

$$\mathbf{P}(A|w) = \frac{\mathbf{P}(A \cap B_w)}{\mathbf{P}(B_w)}, \tag{2.8}$$

which we will write as

$$\mathbf{P}(A|w) = \frac{\mathbf{P}(A, w)}{\mathbf{P}(w)}. \tag{2.9}$$

Next, we will develop a definition for a conditional future distribution given a history, $\mathbf{P}(\cdot|\mathbf{x}^-)$. We will define $\mathbf{P}(\cdot|\mathbf{x}^-)$ in terms of the conditional distributions $\mathbf{P}(\cdot|w_l)$, where w_l is taken to be the length l suffix of the history $\mathbf{x}^-$. For every word s, we will define

$$\mathbf{P}(s|\mathbf{x}^-) = \lim_{l \to \infty} \mathbf{P}(s|w_l). \tag{2.10}$$

Note that if we let

$$1_A(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in A \\ 0 & \text{otherwise,} \end{cases} \tag{2.11}$$

$\mathbf{P}(A|w_l)$ is the conditional expectation $\mathbf{E}(1_A|\mathbb{H}_l)(\mathbf{x})$ for any sequence $\mathbf{x}$ which matches w. Thus the limit in equation 2.10 may be written $\lim_{l \to \infty} \mathbf{E}(1_A|\mathbb{H}_l)$, which converges almost surely to $\mathbf{E}(1_A|\mathbb{H})(\mathbf{x})$ by theorem B.2.3, which is a martingale convergence theorem. Thus we have

$$\lim_{l \to \infty} \mathbf{P}(\cdot|w_l) = \mathbf{P}(\cdot|\mathbf{x}^-) \tag{2.12}$$

for almost every $\mathbf{x}^-$. This is our definition of $\mathbf{P}(\cdot|\mathbf{x}^-)$.

Our next step is to define the set of histories on which we will condition when defining process states. We have just shown that this definition gives a well-defined conditional future distribution for almost all histories, and we will condition only on those histories.

Let us examine how $\mathbf{P}(\cdot|\mathbf{x}^-)$ can fail to be well-defined for a given history $\mathbf{x}^-$. For each of the countably many words s there may be a set of histories N_s — a set

of measure zero — on which $\lim_{l\to\infty} \mathbf{P}(s|w_l)$ does not converge. For example, if x^- has a suffix w_l with probability zero, then for all $m > l$, $\mathbf{P}(w_m) = 0$, and $\mathbf{P}\big(s|\mathrm{x}^-\big)$ is ill-defined for all s. The union $\mathcal{N}$ of the N_s is itself a null set. On its complement $\mathcal{X}^{\mathbb{Z}} \setminus \mathcal{N}$, which is thus a set of full measure, $\mathbf{P}\big(\cdot|\mathrm{x}^-\big)$ exists. We will call $\mathcal{N}$ the set of *bad histories*, and we will say that a history is *good* if it lies in $\mathcal{X}^{\mathbb{Z}} \setminus \mathcal{N}$.

Now we are ready to define process states.

Definition 2.5.1. A *process state* is a conditional future distribution which arises in conditioning on a non-null history suffix or a good history. That is, a process state is either a conditional future distribution $\mathbf{P}(\cdot|w)$ for some $w \in R$ or it is a conitional future distribution $\mathbf{P}\big(\cdot|\mathrm{x}^-\big)$ for some $\mathrm{x}^- \in \mathcal{X}^- \setminus \mathcal{N}$. Thus the *set S of all process states* for a given process is

$$S = \{\mathbf{P}(\cdot|w)|w \in R\} \cup \big\{\mathbf{P}\big(\cdot|\mathrm{x}^-\big)|\mathrm{x}^- \in \mathcal{X}^- \setminus \mathcal{N}\big\}. \tag{2.13}$$

In section 2.4, we developed the idea of process states in terms of equivalence classes of history suffixes. We have now developed a formal definition of process states, and we have defined a process state to be a conditional future distribution and not an equivalence class of history suffixes. This does not require changing how we think about process states, because there is a natural correspondence between equivalence classes as we described them in section 2.4 and conditional future distributions induced by non-null history suffixes. Recall that two history suffixes are said to be equivalent if they induce the same conditional future distribution. Thus every equivalence class has an associated conditional future distribution. At the same time, every process state is induced by at least one non-null history suffix or at least one good history. All history suffixes which induce a given process state are automatically equivalent to each other, as are all histories which induce a given process state. So every process state has an associated equivalence class of history suffixes, an equivalence class of histories, or both.

In addition, some of our terminology will refer to process states as if they were equivalence classes of history suffixes. In particular, we will say that a history or history suffix is a *member* or an *element* of the process state which it *induces*. In addition, we define the map $G : \mathcal{X}^- \to S$ which takes a non-null history to the process state it induces, as $G\big(\mathrm{x}^-\big) = \mathbf{P}\big(\cdot|\mathrm{x}^-\big)$

Definition 2.5.2. The *inducing set of histories* of a process state $\mathbf{A}$ is the set $G^{-1}(\mathbf{A})$ of all histories which induce $\mathbf{A}$.

$G^{-1}(\mathbf{A})$ may be empty; this occurs if $\mathbf{A}$ is induced only by history suffixes.

We will want to define a probability to $G^{-1}(\mathbf{A})$. As it happens, we do not know that $G^{-1}(\mathbf{A}) \times \mathcal{X}^+$ — which lies in $\mathcal{X}^{\mathbb{Z}}$ — is measurable. This next result tells us that, if it does not lie in the σ-field $\mathbb{H}$, $G^{-1}(\mathbf{A}) \times \mathcal{X}^+$ differs from a measurable set by a subset of a set of measure zero. Essentially, this means the we can reasonably assign a measure to it. From here on, we will do so without comment.

Proposition 2.5.3. For any process state $\mathbf{A}$, $G^{-1}(\mathbf{A}) \times \mathcal{X}^+$ can be written as the intersection of a measurable set and a set of full measure.

Proof. In this proof, we will be referring to truncations of both the history and the future. We will use l for history length and k for future length. Also, the reader is reminded that the process state $\mathbf{A}$ is a probability distribution on the future space, and so it makes sense to refer to $\mathbf{A}(w)$, the probability $\mathbf{A}$ assigns to a word w.

For every natural number l and word w, and for every history $\mathbf{x}^-$, define

$$f_l^w(\mathbf{x}^-) = \mathbf{E}(1_{A_w}|\mathbb{H}_l)(\mathbf{x}^-) = \mathbf{P}(w|x_{-l}\dots x_{-1}) \tag{2.14}$$

and

$$f^w(\mathbf{x}^-) = \mathbf{E}(1_{A_w}|\mathbb{H})(\mathbf{x}^-) = \mathbf{P}(w|\mathbf{x}^-). \tag{2.15}$$

By B.2.4, we know that $\lim_{l\to\infty} f_l^w = f^w$ almost surely. f_l^w is a measurable function, and so its limit f^w must be measurable.

If we have two process states $\mathbf{B}$ and $\mathbf{C}$, we will say that they are *k-equivalent* if, for all words w of length less than or equal to k, $\mathbf{B}(w) = \mathbf{C}(w)$. Let $\mathbf{A}$ be a process state and let $A_k(\mathbf{A})$ be the set of all histories $\mathbf{x}^-$ which induce process states $G(\mathbf{x}^-)$ that are k-equivalent to $\mathbf{A}$. In other language, that is,

$$A_k(\mathbf{A}) = \{\mathbf{x}^- | \text{ for all } w \text{ with } |w| \le k,\ \mathbf{P}(w|\mathbf{x}^-) = \mathbf{A}(w)\}. \tag{2.16}$$

That is,

$$\begin{aligned} A_k(\mathbf{A}) &= \bigcap_{w:\,|w|=k} \{\mathbf{x}^- | \mathbf{P}(w|\mathbf{x}^-) = \mathbf{A}(w)\} \\ &= \bigcap_{w:\,|w|=k} (f^w)^{-1}(\mathbf{A}(w)). \end{aligned} \tag{2.17}$$

$\{\mathbf{A}(w)\}$ is a measurable set and f^w is a measurable function, so the inverse image $(f^w)^{-1}(\mathbf{A})$ must be measurable. Thus $A_k(\mathbf{A})$ is a finite intersection of measurable set and must be measurable. Further, if we take

$$A(\mathbf{A}) = \bigcap_{k=1}^{\infty} A_k(\mathbf{A}), \tag{2.18}$$

$A(\mathbf{A})$ is a countable intersection of measurable sets, and so it is measurable.

Now, every history in $A(\mathbf{A}) \cap \left(\mathcal{X}^{\mathbb{Z}} \setminus \mathcal{N}\right)$ induces the process state $\mathbf{A}$, and no history which is not in $A(\mathbf{A})$ induces $\mathbf{A}$. Thus $A(\mathbf{A}) \cap \left(\mathcal{X}^{\mathbb{Z}} \setminus \mathcal{N}\right) = G^{-1}(\mathbf{A})$. Note that the $\mathcal{X}^{\mathbb{Z}} \setminus \mathcal{N}$ has full measure, so we are done. ■

The processes addressed in this section and the previous one output bi-infinite sequences, but the idea of a process state is no less relevant to processes with finite or semi-infinite output. The essential idea is that a state is a prediction of, or distribution on, the future reached by some knowledge of the past.

Before going on to the next section, let us look an example. This process has elements of both of the examples from the previous section. The alphabet is $\mathcal{X} = \{0, 1\}$. We first define a nonstationary probability measure $\mathbf{Q}$ on $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}\right)$, which generates sequences $\mathbf{y} = (\ldots y_{-1} y_0 y_{1\ldots}) \in \mathcal{X}^{\mathbb{Z}}$ such that

1. if t is even, $y_t = 1$, and
2. if t is odd, then y_t is either 0 or 1 with probability $\frac{1}{2}$ each.

Now, we define the process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ by $\mathbf{P} = \frac{1}{2}(\mathbf{Q} + T(\mathbf{Q}))$, where T is the shift map. In words, our process consists of alternating 1s and coin flips, and the coin flips are equally likely in either even or odd positions. We will not prove that this is a process, nor will we do so for the examples in the next section. (In chapter 3, we will develop a systematic approach to calculating probabilities of words, which will make it practical to present such proofs). The process $\mathcal{P}$ is called the *band-merging process* for historical reasons[15].

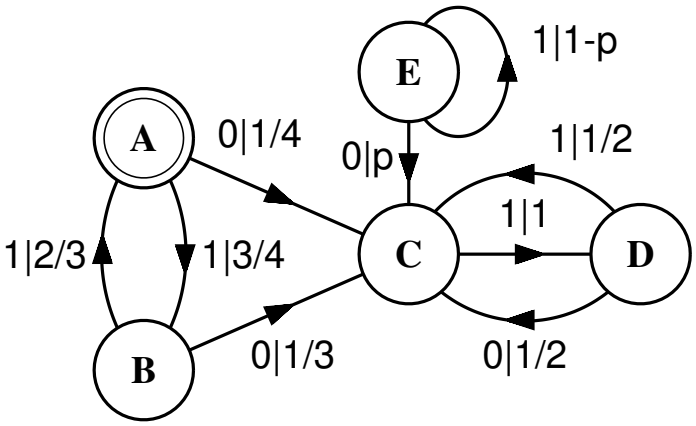

Fig. 2.1 Process state graph representation of the band-merging process. **A**, **B**, **C**, **D**, and **E** are the process states. **A** is the start state (induced by λ). The labels on the edges consist of a symbol followed by a vertical bar and a transition probability. Here p= $1 - \frac{\sqrt{2}}{2}$.

Figure 2.1 is a process state graph representation of the band-merging process. The small circles ($\mathbf{A}$, $\mathbf{B}$, C, $\mathbf{D}$, and $\mathbf{E}$) represent process states. The double circle for state $\mathbf{A}$ indicates that it is the process state induced by λ, which is called the *start state*. The process is said to be *in* a state if the known history or history suffix induces that state. The edges represent possible transitions between states and the labels are of the form *symbol|transition probability*. For instance, the edge from state $\mathbf{A}$ to state $\mathbf{B}$, labeled with "1|3/4" indicates that if the process is in state $\mathbf{A}$, then with probability $\frac{3}{4}$ it will emit a 1, after which it will be in state $\mathbf{B}$.

The band-merging process has five process states. State $\mathbf{A}$ is induced by any history suffix consisting of an even number of 1s, and state $\mathbf{B}$ is induced by any history suffix consisting of an odd number of 1s. States C and $\mathbf{D}$ correspond to any history or history suffix ending with a 0 followed by an even or odd number of 1s, respectively, and state $\mathbf{E}$ is induced only by the history consisting entirely of 1s. State $\mathbf{E}$ is an interesting example of a state which we could ignore because it is irrelevant to the study of the process — it induced only by a single history which has mass zero — but which is well defined. Such states are said to be *elusive*. We will discuss elusive states more in the next section, after which we will usually ignore them.

2.6 Transient and Recurrent States

In section 2.5, we defined a process state to be a distribution on the future which

is induced by a finite length history suffix or a semi-infinite history. This means that there may be some process states which are induced by history suffixes and not by any histories, some states which are induced by histories and not by history suffixes, and some states that are induced by histories and history suffixes. In this section we classify process states by the history objects which induce them.

First, we will define terms for the above distinctions.

Definition 2.6.1. A process state is *infinitely preceded* if it is induced by at least one good history.

Definition 2.6.2. A process state is *reachable* if it is induced by at least one history suffix w with $\mathbf{P}(w) > 0$.

In addition, we need to define one more property. For a process state $\mathbf{A} \in S$, consider $G^{-1}(\mathbf{A})$, the set of histories which induce $\mathbf{A}$. We know that we can assign this set a measure, for which we will now use the shorthand $\mathbf{P}(\mathbf{A}) = \mathbf{P}\big(G^{-1}(\mathbf{A})\big)$, where $\mathbf{P}^-$ is the measure on the history space. That is, if we have seen a history s, $\mathbf{P}(\mathbf{A})$ is the probability that s induces process state $\mathbf{A}$.

Definition 2.6.3. A process state is *recurrent* if $\mathbf{P}(\mathbf{A}) > 0$.

The term *positive recurrent* is often used for this concept [13]. We have chosen to use *recurrent* for brevity and because null recurrence does not happen in the systems of interest here.

Now, for any process state, we may ask three questions. Is it reachable? Is it infinitely preceded? And, is it recurrent? There are eight triples of answers, of which three are impossible and five are observed.

Proposition 2.6.4.

1. Every unreachable process state is infinitely preceded.
2. Every recurrent process state is infinitely preceded.

This proposition asserts that the following three kinds of states do not occur:

- unreachable recurrent states which are not infinitely preceded,
- reachable recurrent states which are not infinitely preceded, and

- unreachable states which are neither recurrent nor infinitely preceded.

Proof. Statement 1 is automatic from the definition of a process state, since every process state is induced by at least one history or history suffix. Statement 2 follows from definitions 2.6.1 and 2.6.3. Let $\mathbf{A}$ be a recurrent process state. Then $G^{-1}(\mathbf{A})$ has positive measure and thus is nonempty. Since $G^{-1}(\mathbf{A})$ contains only good histories, then it contains at least one good history which induces $\mathbf{A}$. ■

Definition 2.6.5. A process state is *transient* if it is reachable and not recurrent. A transient state is said to be *strictly transient* if it is not infinitely preceded.

Definition 2.6.6. A process state is *elusive* if it is unreachable and not recurrent.

Unlike the other kinds of states we have discussed, elusive states can often be ignored. Whereas every reachable state can be induced by at least one word of positive probability and a recurrent state can be induced by a set of histories of positive probability, an elusive state can only be induced by events of zero probability. Thus any countable set of elusive states can be ignored without changing the process' probability measure. We will usually omit elusive states for brevity. However, in some processes the existence and structure of the elusive states is implied by the recurrent states and in others the set of all elusive states is uncountable and has positive measure, so we cannot forget them completely.

Proposition 2.6.4 established that there are five classes of states. This, together with definitions 2.6.5 and 2.6.6, gives us names for them. Table 2.1 summarizes the relevant information about each type. In figure 2.2, we have examples of all types except unreachable recurrent states. In order to present processes which have such states in a reasonable form, we will need to develope more machinery for presenting processes.

Process state type	Reachable	Infinitely Preceded	Recurrent
Strictly Transient	yes	no	no
Infinitely Preceded Transient	yes	yes	no
Reachable Recurrent	yes	yes	yes
Unreachable Recurrent	no	yes	yes
Elusive	no	yes	no

Table 2.1 Summary of process state types.

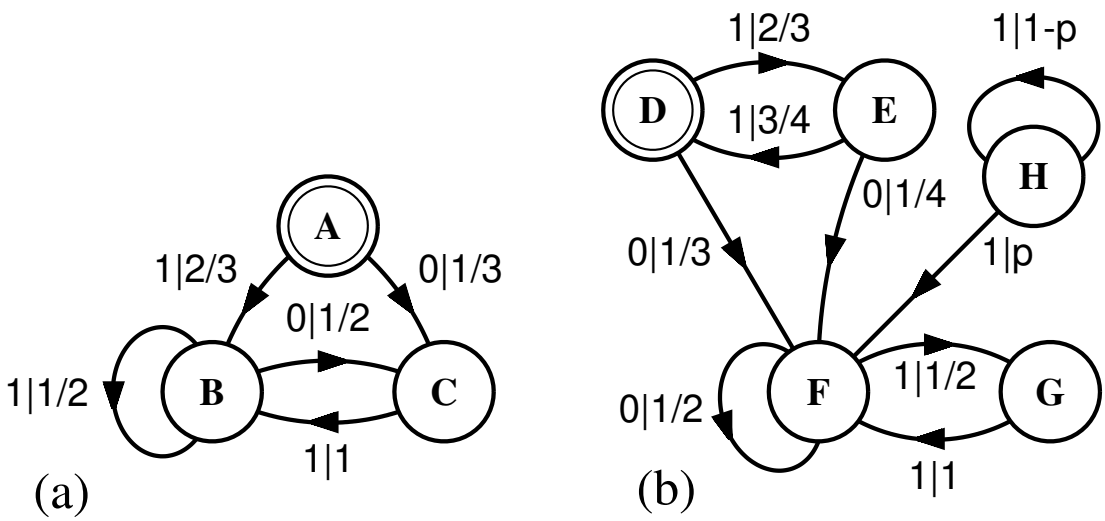

Fig. 2.2 Process state graph representations of two processes. (a) The golden mean process. The process state **A** is reachable and neither infinitely preceded nor recurrent. States **B** and **C** are reachable, infinitely preceded and recurrent. (b) The even process. Here $p = 1 - \frac{\sqrt{2}}{2}$. States **D** and **E** are reachable and infinitely preceded but not recurrent, **F** and **G** are reachable, infinitely preceded and recurrent, and **H** is unreachable, infinitely preceded, and not recurrent. For both of these processes, there may be additional ill-defined states resulting from conditioning on histories in some measure-zero bad set.

2.7 Synchronization

Suppose we are watching the output of the even system, and the history suffix we have seen contains all 1s. This means that our state of knowledge about the future of the process is described by either process state $\mathbf{D}$ or $\mathbf{E}$. It is possible that another observer, who has been watching longer, knows more about the future than we do. In contrast, if the history suffix we have seen contains a 0, then this is not the case. An observer may

have more historical information than we do, but this extra information is irrelevant to the future. In this case we say that we are *synchronized* to the machine.

Definition 2.7.1. A non-null history suffix w is a *synchronizing word* if all good histories and all non-null history suffixes which end in w induce the same process state.

Proposition 2.7.2. If w is a synchronizing word, then w induces a reachable recurrent process state.

Proof. Let $\mathbf{A}$ be the process state induced by w. $\mathbf{A}$ is induced by a history suffix, so it is reachable. Note that w is itself a history suffix that ends in w, so $\mathbf{A}$ is the process state induced by all histories and history suffixes which match w. Further, we know that w is not a null word, and that the set of histories which matches it is a subset of $G^{-1}(\mathbf{A})$. So we have

$$\begin{aligned} \mathbf{P}(\mathbf{A}) &\geq \mathbf{P}(\text{all histories which match } w) \\ &\geq \mathbf{P}(w) > 0. \end{aligned} \tag{2.19}$$

Thus we conclude that $\mathbf{A}$ is recurrent. ■

The converse of Proposition 2.7.2 does not hold. There are process in which nonsynchronizing words induce recurrent states. Also, some processes have no reachable recurrent states, and hence no synchronizing words. These processes either have unreachable recurrent states or have uncountably many elusive states. We will see examples in section 3.5.