3 Hidden Markov Models

In the last chapter, we talked entirely about distributions on sequence space. Although this viewpoint will be necessary for some of our results, processes as we have defined them are not very tractable or structured. A distribution on an infinite set need not lend itself to a finite description, let alone a brief one. In order to do anything concrete, we will need another set of definitions. These are the traditional definitions used in the study of Hidden Markov Models [3,16,17]. In this chapter, we define Hidden Markov Models and then study how they represent processes. We will look at how to represent the process states of a process defined by an HMM. And we will conclude the chapter with a result on the structure of these processes' sets of process states.

The material in sections 3.1, 3.2, and 3.3 is fairly standard in the literature on HMMs, appearing in such works as [3,16]. The contents of section 3.4 has probably all been deduced before. Mixed states, for example, appear in [1], although not with that name. What is new is stating this material in terms of process states. And the material in section 3.6 is entirely the author's though some similar results are known.

We will be careful to keep clear the distinction between the process and the way it's presented to us. By *process*, we will always mean a stationary distribution on a sequence space. When we refer to a finite specification of a process, such as a Hidden Markov Model, we will call it a presentation of a process, or simply a *presentation*.

3.1 Notation for Markov Chains

Before we start on Hidden Markov Models, we will define a Markov Chain. This definition and the following discussion are not intended to be complete; rather, they are intended to introduce the reader to the notation we will be using.

Definition 3.1.1. An n-state *Markov Chain* (MC) is a triple (V, P, π), where V is a finite set of size n, P is an $n \times n$ matrix, and π is a length n row vector, such that

(i) Each row of P has sum one,
(ii) $\sum_i \pi_i = 1$, and
(iii) $\pi P = \pi$.

Elements of V are called states, P is called the transition matrix, and π is called a stationary distribution over the states V.

Note that this definition requires a Markov Chain to have finitely many states. At times in the following, we will discuss both countably and uncountably infinite state Markov Chains, but we will not define them rigorously.

If we let $\mathbb{V}$ be the σ-field defined by the cylinder sets on $V^{\mathbb{Z}}$, then $\left(V^{\mathbb{Z}}, \mathbb{V}\right)$ is a measurable space. We define a distribution $\overline{\mathbf{P}}$ as follows: if $v = v_0 v_1 \dots v_{l-1}$, with all $v_i \in V$, we define

$$\overline{\mathbf{P}}(v) = \pi_{v_0} P_{v_0 v_1} \dots P_{v_{l-2} v_{l-1}}, \tag{3.1}$$

and we define $\overline{\mathbf{P}}(\lambda) = 1$. Equation 2.2 is satisfied trivially. We will verify equation 2.3 and invoke theorem B.1.1 to show that $\overline{\mathbf{P}}$ is a stationary probability distribution. If $z \in V$, then

$$\overline{\mathbf{P}}(\lambda z) = \overline{\mathbf{P}}(z\lambda) = \overline{\mathbf{P}}(z) = \pi_z. \tag{3.2}$$

Thus,

$$\sum_{z \in V} \overline{\mathbf{P}}(z\lambda) = \sum_{z \in V} \overline{\mathbf{P}}(\lambda z) = \sum_{z \in V} \pi_z = 1 = \overline{\mathbf{P}}(\lambda). \tag{3.3}$$

If $v = v_0 \dots v_{l-1}$, we have$\overline{\mathbf{P}}(vz) = \overline{\mathbf{P}}(v) P_{v_{l-1} z}$. Thus

$$\sum_{z \in V} \overline{\mathbf{P}}(vz) = \overline{\mathbf{P}}(v) \sum_{z \in V} P_{v_{l-1} z} = \overline{\mathbf{P}}(v) \cdot 1 = \overline{\mathbf{P}}(v). \tag{3.4}$$

On the other hand,

$$\overline{\mathbf{P}}(zv) = \pi_z P_{z v_0} P_{v_0 v_1} \dots P_{v_{l-2} v_{l-1}}, \tag{3.5}$$

so

$$\sum_{z \in V} \overline{\mathbf{P}}(zv) = \left(\sum_{z \in V} \pi_z P_{z v_0} \right) P_{v_0 v_1} \dots P_{v_{l-2} v_{l-1}}. \tag{3.6}$$

But $\sum\limits_{z \in V} \pi_z P_{z v_0}$ is the v_0 coordinate of πP and $\pi P = \pi$, so

$$\sum_{z \in V} \pi_z P_{z v_0} = \pi_{v_0} \tag{3.7}$$

and

$$\sum_{z \in V} \overline{\mathbf{P}}(zv) = \overline{\mathbf{P}}(v). \tag{3.8}$$

Thus the Markov Chain (V, P, π) defines a process $\left(V^{\mathbb{Z}}, \mathbb{V}, \overline{\mathbf{P}}\right)$.

We conclude this section with at definition which we will need in section 3.2.

Definition 3.1.2. If (V, P, π) is a Markov Chain and $C \subset V$, we say that C is a *recurrent component* of the Markov Chain if:

(i) For all $u, v \in C$, there exists an integer $k > 0$ such that $P^k_{uv} > 0$, and
(ii) For all $u \in C$, for all $v \in V \setminus C$, and for all integers $k > 0$, we have $P^k_{uv} = 0$.

Here, P^k means the kth power of the matrix P.

Definition 3.1.3. A finite Markov Chain is *reducible* if it has more than one recurrent component.

If a Markov Chain is reducible, it is often appropriate to think of it as two or more separate Markov Chains. A Markov Chain which has exactly one recurrent component is said to be *irreducible*.

3.2 Hidden Markov Models

In this section, we will give a definition of a Hidden Markov Model (HMM), and we will show how an HMM specifies a process.

A Hidden Markov Model is a Markov Chain with an associated output mechanism which takes either states or transitions between states to either symbols or distributions on symbols. We will refer to the Markov Chain as the *underlying Markov Chain* of the HMM. We will calculate exclusively with finite presentations — those in which the Markov Chain has finitely many states. However, we will, at times, consider infinite presentations.

Hidden Markov Models appear in the literature in several forms, the most frequent being Functions of a Markov Chain[1] and State-output Hidden Markov Models[16]. These forms are equivalent in the sense that for any HMM in one of these forms, there is an HMM in each of the other forms which defines the same process. The HMMs in this work will be Edge-output Hidden Markov Models, the elements of which are the set of states, the set of symbols, a stationary distribution on those states, and, for each state, a joint distribution on symbols and next states. The following definition formalizes this idea.

Definition 3.2.1. A *Hidden Markov Model (HMM)*is a quadruple $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$, where V and $\mathcal{X}$ are finite sets of sizes $n = |V|$ and $m = |\mathcal{X}|$, $\left\{T^k\right\} = \left\{T^k | k = 0, \ldots, m-1\right\}$ is a set of $n \times n$ matrices, and π is a probability vector with length n. The matrices $\left\{T^k\right\}$ must satisfy

1. For all i such that $0 \leq i \leq n-1$

$$\sum_{j,k} T^i_{ij} = 1, \tag{3.9}$$

2. and for all i,j such that $0 \leq i, j \leq n-1$ and $0 \leq k \leq m-1$,

$$T^k_{ij} \geq 0. \tag{3.10}$$

Finally, π must satisfy

$$\pi_j = \sum_{i,k} \pi_i T^k_{ij}. \tag{3.11}$$

The *underlying Markov Chain* of a Hidden Markov Model is a the Markov Chain $\left(V, \sum_k T^k, \pi\right)$.

Elements of V, called *presentation states*, are the states of the underlying Markov Chain. Elements of $\mathcal{X}$ are called *symbols*, as in chapter 2. Unless we have reason to do otherwise, we will use $V = \{0, 1, \ldots, n-1\}$ or $V = \{A, B, \ldots\}$ and $\mathcal{X} = \{0, 1, \ldots, m-1\}$. The $\left\{T^k\right\}$, called the *joint matrices*, represent a set of joint distributions on next states $j \in V$ and output symbols $k \in \mathcal{X}$ in the following way. If $i, j \in V$ and $k \in \mathcal{X}$ and the Markov Chain is in state i, then the probability that the next symbol emitted will be k and the next state will be j is

$$\mathbf{P}(j, k|i) = T^k_{i,j}. \tag{3.12}$$

The last element of the quadruple is π, which is a *stationary distribution*. Most definitions of HMMs found in the literature have an initial distribution instead of a stationary distribution. The difference is that an initial distribution may be any distribution over the states, whereas the stationary distribution is constrained to satisfy equation 3.11. Using a stationary distribution here makes the resulting process stationary.

If the underlying Markov Chain has a single recurrent component, then π is uniquely determined by the joint matrices. If, however, the underlying Markov Chain has more

than one recurrent component, then π is only partially determined. Choosing a stationary distribution is then tatamount to choosing a distribution over the components.

In addition, we will define a few auxiliary matrices. The *transition matrix* P of a Hidden Markov Model is defined by

$$P_{ij} = \sum_k T_{ij}^k. \tag{3.13}$$

The output matrix B is an $n \times m$ matrix such that B_{jk} gives the probability of emitting the symbol $k \in \mathcal{X}$ while in the state $j \in V$. B is define by

$$B_{ki} = \sum_j T_{ij}^k. \tag{3.14}$$

The conditions imposed on the joint matrices ensure that P and B are stochastic matrices, that is, their rows sums are all equal to 1. Also, we have $\pi P = \pi$, and we can write the underlying Markov Chain of the HMM as (V, P, π).

A warning to readers familiar with state-output Hidden Markov Models defined in terms of transition and output matrices — our choice of notation may be misleading to your intuition. The auxiliary matrices P and B are **not** always sufficient to recover the joint matrices $\{T^k\}$. For example, if we start with a state-output HMM, the joint matrices can be constructed as $T_{ij}^k = P_{ij}B_{jk}$, and equations 3.13 and 3.14 will be satisfied. That is, if we compute the right hand sides of 3.13 and 3.14, we will recover our original transition and output matrices. But, if we start with a set of joint matrices, compute the transition and output matrices by equations 3.13 and 3.14, and then compute $P_{ij}B_{jk}$, the result need not be the joint matrices. Doubtful readers are encouraged to perform the calculations themselves on the two-state, two symbol process with joint matrices $T^0 = \begin{pmatrix} 0 & 0 \\ 1/2 & 0 \end{pmatrix}$ and $T^1 = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1/2 \end{pmatrix}$.

In general, a state-output HMM may be built from an edge-output HMM, but the state-output HMM may need to have a greater number of states, because edge-output HMMs have more degrees of freedom per state than state-output HMMs. Given an edge-output HMM $(V, \mathcal{X}, \{T^k\}, \pi)$, we can construct an equivalent state-output HMM with set of states $U = V \times V$ as follows: if $a, b, c, d \in V$, then we have $(a, b), (c, d) \in U$. Let

$$P_{(a,b),(c,d)} = \begin{cases} \sum\limits_{k \in \mathcal{X}} T_{cd}^k & b = c \\ 0 & b \neq c, \end{cases} \tag{3.15}$$

and let

$$B_{(a,b),k} = \frac{T^k_{a,b}}{\sum\limits_{l \in \mathcal{X}} T^l_{a,b}}. \tag{3.16}$$

3.3 HMMs as Processes

An HMM presentation defines a process. That is, $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ determines a probability distribution $\mathbf{P}$ and thus a process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$. Let us see how this works.

First, we suppose that the presentation's underlying Markov Chain is in the state $i \in V$. Let k be a symbol and $j \in V$ be a presentation state. We want to know $\mathbf{P}(k|i)$, the probability that the next symbol will be k, and $\mathbf{P}(j|i,k)$, the probability that next presentation state will be j if the next symbol is k. These are straightforward to calculate from the presentation.

$$\mathbf{P}(k|i) = \sum_j \mathbf{P}(j,k|i) = \sum_j T^k_{ij} \tag{3.17}$$

$$\mathbf{P}(j|i,k) = \frac{\mathbf{P}(j,k|i)}{\mathbf{P}(k|i)} = \frac{T^k_{ij}}{\sum\limits_l T^k_{il}} \tag{3.18}$$

Next, instead of assuming that the current presentation state is i, that is, $\mathbf{P}(i) = 1$, we assume that it has distribution μ. To calculate the analogous quantities, $\mathbf{P}(k|\mu)$ and $\mathbf{P}(j|k,\mu)$, we start by calculating $\mathbf{P}(j,k|\mu)$. After that, the answers are essentially the same as above.

$$\mathbf{P}(j,k|\mu) = \sum_i \mu_i \mathbf{P}(j,k|i) = \left(\mu T^k\right)_j \tag{3.19}$$

$$\mathbf{P}(k|\mu) = \sum_j \mathbf{P}(j,k|\mu) = \sum_j \left(\mu T^k\right)_j \tag{3.20}$$

$$\mathbf{P}(j|\mu,k) = \frac{\mathbf{P}(j,k|\mu)}{\mathbf{P}(k|\mu)} = \frac{\left(\mu T^k\right)_j}{\sum\limits_j \left(\mu T^k\right)_j} \tag{3.21}$$

If we denote the column vector $(1,\ldots,1)^T$ by $\vec{1}$, we can write $\mathbf{P}(k|\mu) = \mu T^k \vec{1}$.

Now, define a map C_k, which takes distributions μ on the states V to distributions on V, by

$$C_k(\mu) = \mu T^k / \mu T^k \vec{1}. \tag{3.22}$$

We then have $\mathbf{P}(j|\mu, k) = (C_k(\mu))_j$. We think of μ as representing our state of knowledge about the internal state of the process. The C_k should be thought of as update maps: they take a distribution μ at one time and update it to reflect the passage of time and the latest observation k.

Having addressed single symbols, we are ready to address words. We begin with a word w of length two, $w = w_0 w_1$. $\mathbf{P}(w|\mu)$ factors to $\mathbf{P}(w_0|\mu) \cdot \mathbf{P}(w_1|w_0, \mu)$. The first of these terms is a case we have just treated in 3.20. For the second, if we update μ to $C_{w_0}(\mu)$, it reduces to the same case: $\mathbf{P}(w_1|w_0, \mu) = \mathbf{P}(w_1|C_{w_0}(\mu))$. We now expand and simplify,

$$\begin{aligned}
\mathbf{P}(w|\mu) &= \mathbf{P}(w_0|\mu) \cdot \mathbf{P}(w_1|C_{w_0}(\mu)) \\
&= \left(\mu T^{w_0} \vec{1}\right)\left(C_{w_0}(\mu) T^{w_1} \vec{1}\right) \\
&= \left(\mu T^{w_0} \vec{1}\right)\left(\frac{\mu T^{w_0}}{\mu T^{w_0} \vec{1}}\right) \cdot \left(T^{w_1} \vec{1}\right) \\
&= \mu T^{w_0} T^{w_1} \vec{1}
\end{aligned} \tag{3.23}$$

By similar manipulations, we have

$$\begin{aligned}
(C_{w_1} \circ C_{w_0})(\mu) &= C_{w_1}\left(\frac{\mu T^{w_0}}{\mu T^{w_0} \vec{1}}\right) \\
&= \frac{\left(\mu T^{w_0} / \mu T^{w_0} \vec{1}\right) T^{w_1}}{\left(\mu T^{w_0} / \mu T^{w_0} \vec{1}\right) T^{w_1} \vec{1}} \\
&= \frac{\mu T^{w_0} T^{w_1}}{\mu T^{w_0} T^{w_1} \vec{1}}
\end{aligned} \tag{3.24}$$

This extends to words of arbitrary length. If w is a word of length l, then $\mathbf{P}(w|\mu) = \mu T^{w_0} T^{w_1} \ldots T^{w_{l-1}} \vec{1}$ and the updated distribution over the presentation states is

$$\left(C_{w_{l-1}} \circ \ldots \circ C_{w_0} \ldots\right)(\mu) = \frac{\mu T^{w_0} \ldots T^{w_{l-1}}}{\mu T^{w_0} \ldots T^{w_{l-1}} \vec{1}} \tag{3.25}$$

Now, if we use the stationary distribution π in place of the arbitrary distribution μ, we have a stationary process.

Lemma 3.3.1. There is a unique stationary process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ such that for all words $w = w_0 \ldots w_{l-1}$,

$$\mathbf{P}(w) = \pi T^{w_0} \ldots T^{w_{l-1}} \vec{1}. \tag{3.26}$$

Proof. We will simply verify equations 2.2 and 2.3 and invoke theorem B.1.1. First, $\mathbf{P}(\lambda) = \pi \vec{1} = \sum_i \pi_i = 1$. This takes care of 2.2. Second, for $z \in \mathcal{X}$,

$$\mathbf{P}(wz) = \pi T^{w_0} \ldots T^{w_{l-1}} T^z \vec{1}. \tag{3.27}$$

Thus,

$$\sum_{z \in \mathcal{X}} \mathbf{P}(wz) = \pi T^{w_0} \ldots T^{w_{l-1}} \left(\sum_{z \in \mathcal{X}} T^z \right) \vec{1}. \tag{3.28}$$

But the rows of $\sum_z T^z$ sum to one, so $\left(\sum_z T^z \right) \vec{1} = \vec{1}$. Hence,

$$\sum_{z \in \mathcal{X}} \mathbf{P}(wz) = \pi T^{w_0} \ldots T^{w_{l-1}} \vec{1} = \mathbf{P}(w). \tag{3.29}$$

Similarly,

$$\sum_{z \in \mathcal{X}} \mathbf{P}(zw) = \pi \left(\sum_{z \in \mathcal{X}} T^z \right) T^{w_0} \ldots T^{w_{l-1}} \vec{1}. \tag{3.30}$$

But $\pi \left(\sum_z T^z \right) = \pi$, so

$$\sum_{z \in \mathcal{X}} \mathbf{P}(zw) = \pi T^{w_0} \ldots T^{w_{l-1}} \vec{1} = \mathbf{P}(w). \tag{3.31}$$

Thus the hypotheses of theorem B.1.1 are satisfied.■

Definition 3.3.2. The *process defined by an HMM presentation* $\left(V, \mathcal{X}, \{T^k\}, \pi\right)$ is the process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ which assigns the probability $\mathbf{P}(w) = \mathbf{P}(w|\pi)$ for any word w of symbols in $\mathcal{X}$.

Over the course of this dissertation, we will be doing many calculations containing expressions of the form $T^{w_0} \ldots T^{w_{l-1}}$. In order to shorten these expressions, we will define the matrix T^w for any word w. If $w = w_0 \ldots w_{l-1}$, then we define $T^w = T^{w_0} \ldots T^{w_{l-1}}$. For the empty word λ, we define $T^\lambda = I$. Thus, for any pair of words, w and z, we have $T^{wz} = T^w T^z$. In this notation, the probability of a word w is $\mathbf{P}(w) = \pi T^w \vec{1}$.

As we have seen, matrix presentations are convenient for calculation. Intuitive interpretation, on the other hand, is often easier with some other forms of presentation. For this reason, we will introduce a new form of presentation, which we will call a *labeled directed graph*. Examples of labeled directed graph presentations may be found in section 3.5. It is worth noting that, while labeled directed graph presentations are often quite clear, they become less intelligible as the number of edges per state increases. For example, compare figures 3.3 and 3.6 on pages 44 and 46.

We have already seen process state graph presentations in sections 2.5 and 2.6; the presentations we define here are related, but distinct. Here the nodes of a labeled directed graph represent presentation states, and not process states as was the case before. Process state graphs are deterministic — that is, they cannot have two or more edges leaving the same state labeled with the same symbol. Labeled directed graphs do not have this restriction.

A labeled directed graph is a directed graph in which the nodes represent presentation states and the edges represent possible transitions. Each edge is labeled with a symbol and a probability. An edge from state i to state j which is labeled with $k|p$ corresponds to an entry in a joint matrix: $T_{ij}^k = p$. That is, k is a symbol and p is a probability, and whenever the labeled directed graph is in state i, it has probability p of following this edge, and if it does so it will output a k and go to state j. We can translate an HMM into a labeled directed graph by drawing a node for each state of the HMM and an edge for each nonzero T_{ij}^k. Similarly, we can usually translate a labeled directed graph into an HMM. We let V be the set of nodes in the graph and $\mathcal{X}$ be the set of all symbols which are appear on the edges of the graph. For each $i, j \in V$ and $k \in \mathcal{X}$, if there is an edge from state i to state j which is labeled with $k|p$ for some p, then we set $T_{ij}^k = p$, and otherwise we set $T_{ij}^k = 0$. The one piece of an HMM which is not present in a labeled directed graph is the stationary distribution π. If there is only one possible stationary distribution for the set of joint matrices, then the labeled directed graph is a complete presentation, and it defines a process. If there is more than one — that is, if the underlying Markov Chain has several recurrent components or is periodic[13] — then the labeled directed graph does not specify a process.

A given process may have many presentations, and determining whether or not two presentations describe the same process is nontrivial [7,2]. For example, the

$$V = \{0,1,2\},\ \mathcal{X} = \{0,1\},\ \pi_A = (\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2})$$
$$T^0 = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},\ T^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

Fig. 3.1 Process "simple nondeterministic source," presentation A.

$$W = \{0,1\},\ \mathcal{X} = \{0,1\},\ \pi_B = (\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2})$$
$$U^0 = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & 0 \end{pmatrix},\ U^1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix}$$

Fig. 3.2 Process "simple nondeterministic source," presentation B.

presentations in figures 3.1 and 3.2 define the same process. To show that presentations A and B are equivalent, it is sufficient to show that, for every finite word w, $\pi_A T^w \vec{1} = \pi_B U^w \vec{1}$. In this case, it can be done by induction. However such proofs are at best computationally messy and are not very illuminating. In section 4.3, we will develop a systematic approach to equivalence of presentations. We will prove that A and B are equivalent there.

Not all processes can be presented as finite HMMs. For example, consider the *modified nested parentheses process*[18], a process with the alphabet of `(` , `)`, and `!`. (The term *modified* refers to the presence of the `!` symbol.) One way to represent this process is as a single presentation state and a counter which holds a nonnegative integer. If the counter is set to zero, then with probability $\frac{1}{3}$, the machine outputs a `(` and sets the counter to one, and with probability $\frac{2}{3}$ it outputs a `!` and leaves the counter at zero. If the counter is not set to zero, then with probability $\frac{2}{3}$ the machine outputs a `)` and decrements the counter and with probability $\frac{1}{3}$ it outputs a `(` and increments the counter. If the initial value of the counter is drawn from the appropriate distribution, this description defines a (stationary) process. This process always outputs balanced strings of parentheses between any consecutive pair of `!` symbols, and there is no upper bound to the number of levels of nesting. We will prove in section 3.6 that there is no HMM presentation for this process.

Simply stated, in this section we have shown how to get a process from an HMM.

But consider the inverse problem — suppose we have a process, and we want an HMM presentation for it. Because a process can have more than one HMM presentation, we cannot expect a unique answer. And, as the modified nested parentheses process illustrates, we cannot always expect any answer at all. This is a form of the problem of HMM reconstruction, and nothing we have seen here so far suggests a way of approaching it.

Finally, we can define the class of processes which are the subject of this dissertation, stochastic finite automata. A *stochastic finite automaton* (SFA) is a process which has a finite HMM presentation. In section 3.6, we will give a necessary condition for a process to be an SFA. Notably, this condition will, among other things, suggest an approach to HMM reconstruction.

3.4 Mixed States

In section 2.5, we defined process states in rather abstract terms, and in section 3.2 we described HMMs in more concrete terms. In this section, we will bring these threads together and discuss the process states of processes defined by HMM presentations.

Recall that a process state is a conditional future distribution which arises when we condition on a history or a history suffix. Suppose we have a process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ defined by an HMM presentation $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$. What are the process states for this process?

There are some presentations for which the process states coincide with the presentation states. Such presentations are necessarily *deterministic*. This means that, for any given presentation state $i \in V$ and symbol $k \in \mathcal{X}$ there is at most one presentation state $j \in V$ such that the transition from i to j with symbol k is possible, $T^k_{ij} \neq 0$. If a process has a finite deterministic presentation then it is called a *Stochastic Deterministic Finite Automaton* (SDFA). In this case, the presentation states and process states are similar though they may not coincide. SDFAs are an important class of processes; see [19]. However, typical HMMs are not deterministic and the processes they represent are not SDFAs. It is this case which this section addresses.

We will begin with reachable states, those which result from conditioning on a finite history suffix. Suppose s is a history suffix and w is a next word. We have

$$\mathbf{P}(w|s) = \frac{\mathbf{P}(sw)}{\mathbf{P}(s)} = \frac{\pi T^s T^w \vec{1}}{\pi T^s \vec{1}}. \tag{3.34}$$

(If $\mathbf{P}(s) = 0$, then $\mathbf{P}(w|s)$ is not well defined. We will ignore such s throughout this section.) Since the conditional distribution $\mathbf{P}(\cdot|s)$ is the object we are interested in and w is the argument it takes, we will rewrite this as

$$\mathbf{P}(w|s) = \mathbf{P}(\cdot|s)(w) = \frac{\pi T^s}{\pi T^s \vec{1}} T^w \vec{1} \tag{3.35}$$

Here, $\mathbf{P}(\cdot|s)$ shows up as $\pi T^s / \pi T^s \vec{1}$, which is a distribution on the presentation states.

In fact, distributions over the presentation states are close to being process states. If μ is such a distribution, then $\mathbf{P}(\cdot|\mu)$ is the conditional future distribution given the measure μ, defined by $\mathbf{P}(w|\mu) = \mu T^w \vec{1}$. We will show below that all process states can be represented in this way. If two different history suffixes, s and $\overline{s}$, define the same distribution over presentation states — $\pi T^s / \pi T^s \vec{1} = \pi T^{\overline{s}} / \pi T^{\overline{s}} \vec{1}$ — then clearly $\mathbf{P}(\cdot|s) = \mathbf{P}(\cdot|\overline{s})$, so s and $\overline{s}$ lead to the same process state.

Before we proceed, we will introduce a notational convenience. When we have a row vector μ, we often need to *normalize* it, that is, scale it so that the sum of its components is 1. We have been writing the normalization of μ as $\frac{\mu}{\mu\vec{1}}$. We now define N, the *normalizing function*, which takes row vectors to row vectors, by

$$N(\mu) = \frac{\mu}{\mu\vec{1}}. \tag{3.36}$$

With this, we can write $N(\pi T^w)$ instead of $\pi T^w / \pi T^w \vec{1}$.

Definition 3.4.1. A *mixed state* of a presentation is a distribution over the presentation states.

(The name *mixed state* comes from thinking of mixed states as "mixtures" of presentation states. This is similar to the use of "mixed state" in quantum mechanics. It should be noted that Fraser and Dimitriadis have use the term "mixed state" in connection with HMMs to mean something entirely different [12].)

Mixed states are related to process states, but they are not quite the same. First, there can be mixed states which do not represent any process states. For example, consider

the process presented by the HMM

$$\begin{aligned} V = \{0,1\}, \mathcal{X} = \{0,1\}, \pi = (\tfrac{1}{2}, \tfrac{1}{2}) \\ T^0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, T^1 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}. \end{aligned} \tag{3.37}$$

This process has only three process states. (If we have seen any history or history suffix of length 1 or more, then we know the entire past and the entire future almost surely — it is either $\ldots 0.10101\ldots$ or $\ldots 1.01010\ldots$. If not, we are conditioning on λ, and we get the futures $10101\ldots$ and $01010\ldots$ with probability $\frac{1}{2}$ each.) The mixed states corresponding to these process states are $(0,1)$, $(1,0)$, and $(\frac{1}{2}, \frac{1}{2})$. The other mixed states do not define process states.

Second, it can happen that two or more different mixed states correspond to a single process state. This can only happen if the presentation in question is not minimal, that is, if it has some redundancy in its states. For example, the process presented by the HMM

$$\begin{aligned} V = \{0,1\}, \mathcal{X} = \{0,1\}, \pi = (\tfrac{1}{2}, \tfrac{1}{2}) \\ T^0 = \begin{pmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 0 \end{pmatrix}, T^1 = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix} \end{aligned} \tag{3.38}$$

is an elaborate presentation of a fair coin, which has only one process state. The mixed states $(1,0)$ and $(0,1)$, which arise as $N(\pi T^0)$ and $N(\pi T^1)$ respectively, represent the same process state.

Definition 3.4.2. Fix a process and an HMM presentation for it. Let $\mathbf{A}$ be a process state and μ a mixed state. If for all next words w we have $\mathbf{A}(w) = \mu T^w \vec{1}$, then we say that μ is a *mixed state version* of the process state $\mathbf{A}$.

Theorem 3.4.3. Suppose we have a process and an HMM presentation for it. Then every process state, except possibly those in a null set, has a mixed state version.

For a reachable process state $\mathbf{A}$, we have essentially already shown this. If s is a history suffix with $\mathbf{P}(s) > 0$ which induces $\mathbf{A}$, $N(\pi T^s)$ is a mixed state version of $\mathbf{A}$. However, for unreachable states, there is no such simple solution. Most of the rest of this section addresses this issue. The proof of this theorem appears on page 41.

To treat this case, we need to work in a probability space which contains both presentation states and symbols. Begin with our HMM $(V, \mathcal{X}, \{T^k\}, \pi)$, and its underlying Markov Chain (V, P, π). These define the observation process $\mathcal{P} = (\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P})$ and the

internal process $\left(V^{\mathbb{Z}}, \mathbb{V}, \overline{\mathbf{P}}\right)$, respectively. We will define the *joint process* of these two to be the process $\mathcal{Q} = \left((V \times \mathcal{X})^{\mathbb{Z}}, \mathbb{J}, \mathbf{Q}\right)$ as follows. The alphabet of the joint processes is $V \times \mathcal{X}$ and thus its sequence space is $(V \times \mathcal{X})^{\mathbb{Z}}$. Its σ-field is the σ-field generated by the cylinder sets in $(V \times \mathcal{X})^{\mathbb{Z}}$. If we have a word $\hat{w} = (v_1, x_1), (v_2, x_2), \ldots, (v_l, x_l)$, we have

$$\mathbf{Q}(\hat{w}) = \pi_{v_1} T^{x_1}_{v_1 v_2} \ldots T^{x_{l-1}}_{v_{l-1} v_l} \left(\sum_{i \in V} T^{x_l}_{v_l i} \right). \tag{3.39}$$

The pair (v, x) corresponds to our original HMM leaving state v and outputting symbol x. Thus $\mathbf{Q}(\hat{w})$ is the probability that the HMM traverses the sequence $v_1, v_2, \ldots, v_l$ of presentation states and, as it does this, emits the word $x_1, x_2, \ldots, x_l$. Specifically, this is the probability that the HMM starts in presentation state v_1, emits x_1 while making a transition to v_2, and then emits x_2 while going to v_3, and so forth. This ends when the HMM emits x_{l-1} during the transition from v_{l-1} to v_l and then emits x_l during a transition to any state. This free choice of the $l+1$th state leads to the sum at the end of equation 3.39. The new process $\mathcal{Q}$ has the HMM presentation $\left(V, V \times \mathcal{X}, \left\{U^k | k \in V \times \mathcal{X}\right\}, \pi\right)$, where if $v \in V$ and $x \in \mathcal{X}$, then $U^{(v,x)}$ is defined by

$$U^{(v,x)}_{ij} = \begin{cases} T^x_{ij} & v = i \\ 0 & v \neq i \end{cases} \tag{3.40}$$

and we can rewrite 3.39 as

$$\mathbf{Q}(\hat{w}) = \pi U^{(v_1, x_1)} \ldots U^{(v_l, x_l)} \vec{1} \tag{3.41}$$

Let $M : V \times \mathcal{X} \to \mathcal{X}$ be the projection map $M(v, x) = x$, and let $M^{\mathbb{Z}} : (V \times \mathcal{X})^{\mathbb{Z}} \to \mathcal{X}^{\mathbb{Z}}$ be the projection map on sequence spaces which applies M at each time index: $M^{\mathbb{Z}}(\ldots z_i z_{i+1} \ldots) = (\ldots M(z_i) M(z_{i+1}) \ldots)$. Thus for any subsequence $s = s_a s_{a+1} \ldots s_b$, $s_i \in \mathcal{X}$ when we apply M^{-1} to the cylinder set A_s we get the set of all sequences in $(V \times \mathcal{X})^{\mathbb{Z}}$ whose x part matches s,

$$M^{-1}(A_s) = \left\{ z \in (V \times \mathcal{X})^{\mathbb{Z}} | M(z_i) = s_i \text{ for all } i \in a, a+1, \ldots b \right\}. \tag{3.42}$$

It should be clear that M is a measurable function, and that if $A \in \mathbb{X}$, we have $\mathbf{P}(A) = \mathbf{Q}\left(M^{-1}(A)\right)$.

In some sense, defining joint processes is a more natural way of approaching HMMs, than the path we have taken of defining (symbol) processes first and then introducing

HMMs as ways of representing processes. However, the joint process approach leads one's intuition in a direction other than the one in which this work is going. In particular, the joint process approach does not suggest section 3.6, and in fact could lead one to reject it. This is because introducing HMMs and joint processes first puts presentation states in a more fundamental role than process states. In contrast, the insight which led to section 3.6 resulted in part from observing that process states were actually the more fundamental objects. We will use the joint process only in part of this section.

In section 2.5 we defined R to be the set of words $w \in \mathcal{X}^*$ such that $P(w) \neq 0$. We also defined the set of bad histories $\mathcal{N}$ to be the set of all histories, and we showed that $\mathcal{N}$ is a null set. A history $\mathbf{x}^-$ is in $\mathcal{N}$ if $\lim_{l\to\infty} \mathbf{P}(s|w_l)$ does not exist for some $s \in \mathcal{X}^*$, where w_l is the length l suffix of $\mathcal{N}$. In particular, if $\mathbf{x}^-$ is not in $\mathcal{N}$, we know that every suffix w_l of $\mathbf{x}^-$ lies in R.

Definition 3.4.4. If s is either a history suffix in R or a good history, the mixed state $\eta(s)$ is defined to be that mixed state whose ith coordinate satisfies $(\eta(s))_i = \mathbf{Q}(v_0 = i|s)$ for all $i \in V$. We call $\eta(s)$ the mixed state *induced* by s.

How can we compute induced mixed states? If w is a history suffix in R, we can calculate directly, using equation 3.41 and definition 3.4.4. The answer is far less cumbersome than the calculations needed to produce it, and brings us back to the material of pages 36–36.

$$\mathbf{Q}(v_0 = i|w) = \frac{\mathbf{Q}(i,w)}{\mathbf{Q}(w)} = \frac{\mathbf{Q}(v_0 = i, x_{-l}\ldots x_{-1} = w_{-l}\ldots w_{-1})}{\mathbf{Q}(x_{-l}\ldots x_{-1} = w_{-l}\ldots w_{-1})} \tag{3.43}$$

$$\mathbf{Q}(v_0 = i|w) = \frac{\sum_{v_{-l}\ldots v_{-1}\in V}\left(\pi U^{(v_{-l},w_{-l})}\ldots U^{(v_{-1},w_{-1})}\sum_{x\in\mathcal{X}} U^{(i,x)}\vec{1}\right)}{\sum_{v_{-l}\ldots v_{-1}\in V}\left(\pi U^{(v_{-l},w_{-l})}\ldots U^{(v_{-2},w_{-1})}\vec{1}\right)} \tag{3.44}$$

$$\mathbf{Q}(v_0 = i|w) = \frac{\pi\left(\sum_{v_{-l}\in V} U^{(v_{-l},w_{-l})}\right)\ldots\left(\sum_{v_{-1}\in V} U^{(v_{-1},w_{-1})}\right)\sum_{x\in\mathcal{X}} U^{(i,x)}\vec{1}}{\pi\left(\sum_{v_{-l}\in V} U^{(v_{-l},w_{-l})}\right)\ldots\left(\sum_{v_{-1}\in V} U^{(v_{-1},w_{-1})}\right)\vec{1}} \tag{3.45}$$

Note that $\sum_{v\in V} U^{(v,x)} = T^x$ and that $\left(\sum_{x\in\mathcal{X}} U^{(i,x)}\vec{1}\right)_j = \sum_{v\in V, x\in\mathcal{X}} U^{(i,x)}_{j,v} = \delta_{ij}$, which means that $\sum_{x\in\mathcal{X}} U^{(i,x)}\vec{1} = e_i$, the ith standard basis vector. Thus we can write

$$\begin{aligned}\mathbf{Q}(i|w) &= \frac{\pi T^{w_{-l}}\dots T^{w_{-1}}e_i}{\pi T^{w_{-l}}\dots T^{w_{-1}}\vec{1}} \\ &= \frac{\pi T^{w}e_i}{\pi T^{w}\vec{1}} = N(\pi T^w)e_i.\end{aligned} \tag{3.46}$$

Thus, the induced mixed state $\eta(w)$ is simply given by $\eta(w) = N(\pi T^w)$.

Before we address the mixed state $\eta(\mathrm{x}^-)$ induced by a history x^-, we need the following theorem, due to technical difficulties of conditioning on sets of measure zero.

Corollary B.2.4. If $\{\mathcal{F}_n\}$ is an increasing sequence of σ-fields and A is an event, then $\mathbf{P}(A|\mathcal{F}_n) \to \mathbf{P}(A|\mathcal{F})$ almost surely, where $\mathcal{F}$ is the smallest σ-field which contains all of the $\mathcal{F}_n$s.

Proposition 3.4.5. For any history x^-, let s_l denote the length l history suffix $x_{-l}\dots x_{-1}$. For almost every x^-, $\eta(s_l) \to \eta(\mathrm{x}^-)$ as $l \to \infty$.

Proof. For each positive integer l, let $\mathcal{F}_l \subset \mathbb{J}$ be the σ-field generated on $(V\times\mathcal{X})^{\mathbb{Z}}$ by history suffixes $w \in \mathcal{X}^*$ of length l, and let $\mathcal{F}_\infty$ be the σ-field generated by the union of the $\mathcal{F}_l$s. Thus $\mathcal{F}_\infty$ is the set of inverse images under M of sets in the history σ-field $\mathbb{H}$ of the process $\mathcal{P}$. Also, let $A_i \subset (V\times\mathcal{X})^{\mathbb{Z}}$ be the set on which $v_0 = i$. Now, applying theorem B.2.4, we get

$$\mathbf{Q}(A_i|\mathcal{F}_l) \to \mathbf{Q}(A_i|\mathcal{F}_\infty) \tag{3.47}$$

almost surely as $l \to \infty$. For a given history $\mathrm{x}^- \notin \mathcal{N}$, and for each positive integer l, let s_l denote the length l history suffix $x_{-l}\dots x_{-1}$. Now,

$$\mathbf{Q}(A_i|\mathcal{F}_l)(\mathrm{x}^-) = \mathbf{Q}(A_i|s_l) = (\eta(s_l))_i \tag{3.48}$$

since we know that $s_l \in R$. Similarly,

$$\mathbf{Q}(A_i|\mathcal{F}_\infty)(\mathrm{x}^-) = \mathbf{Q}(A_i|\mathrm{x}^-) = \left(\eta(\mathrm{x}^-)\right)_i, \tag{3.49}$$

so equation 3.47 becomes $\eta(s_l) \to \eta(\mathrm{x}^-)$ almost surely as $l \to \infty$, for almost every $\mathrm{x}^- \notin \mathcal{N}$, or simply for almost every x^-.■

The next result establishes that $\eta(s)$ contains all the information about the past which is contained in s and which is relevant to the future.

Proposition 3.4.6 Let w be any word in $\mathcal{X}^*$. If s is a history suffix in R, then $\mathbf{P}(w|s) = \eta(s)T^w\vec{1}$. And if x^- is a good history, then $\mathbf{P}(w|s) = \eta(\mathrm{x}^-)T^w\vec{1}$ almost surely.

Proof. If s is a history suffix, we know that

$$\mathbf{P}(w|s) = \frac{\pi T^s}{\pi T^s\vec{1}}T^w\vec{1} = N(\pi T^s)T^w\vec{1}. \tag{3.50}$$

Since $\eta(s) = N(\pi T^s)$, we have $\mathbf{P}(w|s) = \eta(s)T^w\vec{1}$.

For a good history x^-, let $s_l = x_{-l}\ldots x_{-1}$ for each l, and let $\mathcal{F}_l \subset \mathbb{X}$ be the σ-field generated by the history suffixes of length l. In addition, let $A_w \subset \mathcal{X}^{\mathbb{Z}}$ be the cylinder set of sequences which contain w. Now, if we apply theorem B.2.4, we get $\mathbf{P}(A_w|\mathcal{F}_l) \to \mathbf{P}(A_w|\mathcal{F}_\infty)$ almost surely as $l \to \infty$, or equivalently $\mathbf{P}(w|s_l) \to \mathbf{P}(w|\mathrm{x}^-)$.

On the other hand, we know that $\eta(s_l) \to \eta(\mathrm{x}^-)$ almost surely. The function $\mu \to \mu T^w\vec{1}$ is continuous, so $\eta(s_l)T^w\vec{1} \to \eta(\mathrm{x}^-)T^w\vec{1}$ almost surely. And since $\mathbf{P}(w|s_l) = \eta(s_l)T^w\vec{1}$ almost surely, we know that $\mathbf{P}(w|s_l)$ converges almost surely to both $\mathbf{P}(w|\mathrm{x}^-)$ and to $\eta(\mathrm{x}^-)T^w\vec{1}$, so it must be true that $\eta(\mathrm{x}^-)T^w\vec{1} = \mathbf{P}(w|\mathrm{x}^-)$ almost surely. ■

Proposition 3.4.6 directly implies that the past and the future are conditionally independent given the mixed state induced by the past. At last, we can return to mixed state versions of process states and prove theorem 3.4.3.

Proof of theorem 3.4.3 Let $\mathbf{A}$ be a process state for $\mathcal{P} = (\mathcal{X}, \mathcal{X}^{\mathbb{Z}}, \mathbf{P})$. Then there is either a history or a history suffix which induces $\mathbf{A}$. Let s be any such history or history suffix. For all next words w, $\mathbf{A}(w)$ is defined to be $\mathbf{P}(w|s)$ almost surely, and we know that $\mathbf{P}(w|s) = \eta(s)T^w\vec{1}$, so $\mathbf{A}(w) = \eta(s)T^w\vec{1}$. Thus $\eta(s)$ is a mixed state version of $\mathbf{A}$. ■

Finally, with the remainder of this section, we will define a new presentation, called the *mixed state representation* (MSR). If we start with a presentation $(V, \mathcal{X}, \{T^x\}, \pi)$, let $\overline{V}$ be the set of all mixed states $\eta(s)$ which are induced by a history $s \in R$ or a history suffix $\mathrm{x}^- \notin \mathcal{N}$. Elements of $\overline{V}$ are presentation states of the mixed state representation. That is, presentation states $\overline{V}$ of the MSR are mixed states of the

presentation $(V, \mathcal{X}, \{T^x\}, \pi)$. The mixed state representation is another presentation of the process defined by $(V, \mathcal{X}, \{T^x\}, \pi)$. Notably, it may have infinitely many states. It is with this representation in mind that we use the word *state* in the term *mixed state*.

Suppose what we know of the history of our process $\mathcal{P}$ is that the most recent output word was the history suffix w. Then the next symbol will be $x \in \mathcal{X}$ with probability $\mathbf{P}(x|w) = \eta(w)T^x\vec{1}$, and if x is the next symbol, then the known history word becomes wx. Now, we will look at this transition in terms of the mixed states. Since we know that the history suffix is w, we are in mixed state $\eta(w)$. From $\eta(w)$, the next symbol is x with probability $\mathbf{P}(x|\eta(w)) = \eta(w)T^x\vec{1}$, and if x is chosen as the next symbol, then a transition is made to the MSR state $\eta(wx)$.

In order to use mixed states as states, we need to be able to compute $\eta(wx)$ from $\eta(w)$ without using w. Fortunately, this is not difficult to do.

$$\begin{aligned} \eta(wx) &= N(\pi T^w T^x) \\ &= \frac{\pi T^w T^x}{\pi T^w T^x \vec{1}} \\ &= \frac{\pi T^w T^x / \pi T^w \vec{1}}{\pi T^w T^x 1 / \pi T^w \vec{1}}. \end{aligned} \tag{3.51}$$

Thus we have

$$\eta(wx) = \frac{\eta(w)T^x}{\eta(w)T^x\vec{1}} = N(\eta(w)T^x) = C_x(\eta(w)). \tag{3.52}$$

Note that w does not appear except in $\eta(w)$ and $\eta(wx)$.

Now we can define the mixed state representation. As we have stated, its presentation states are elements of $\overline{V}$, mixed states which are induced by histories or history suffixes. We write them as row vectors $\mu = (\mu_1, \ldots, \mu_{|V|})$. Its symbol set, clearly, will be $\mathcal{X}$. Because $\overline{V}$ may be infinite or even uncountable, we cannot define transition matrices, but we can give equivalent information. Given a state $\mu \in \overline{V}$ and a symbol $x \in \mathcal{X}$, if the current state is μ,

(i) the probability that x is emitted is $\mathbf{P}(x|\mu) = \mu T^x\vec{1}$, and
(ii) if x is emitted, the next state is $C_x(\mu) = N(\mu T^x)$.

We will not address here the issue of whether or not a stationary distribution on $\overline{V}$ exists. To use the mixed state presentation to compute the probability of a word w,

assume presentation starts in state $\pi = \eta(\lambda) \in \overline{V}$ and compute

$$\prod_{i=1}^{l} \mu_{i-1} T^{w_i} \vec{1} \tag{3.53}$$

where $l = |w|$, $\mu_0 = \pi$ and $\mu_i = C_{w_i}(\mu_{i-1}) = N(\mu_{i-1} T^{w_i})$ for $i \geq 1$. It can easily be verified that the result of this calculation is equal to the probability $\mathbf{P}(w) = \pi T^{w_1} T^{w_2} \ldots T^{w_l} \vec{1}$ assigned by $\left(V, \mathcal{X}, \{T^k\}, \pi\right)$. The product in equation 3.53 is simply the product of the quantities which the normalizing function N divides out of the μ_i.

Note that the mixed state representation is deterministic. That is, for any MSR state $\mu \in \overline{V}$ and any symbol $x \in \mathcal{X}$, there is a unique MSR state $C_x(\mu)$ to which a transition involving the emission of x is possible. Further, the MSR states are in one-to-one correspondence with the process states, except perhaps for a set of each of measure zero.

In this section, we defined mixed states and showed that they are intimately related to the process states. In fact, the significance of mixed states is that they give us a way of representing the process states.

3.5 Examples

At this point we will digress from the formal development and present several examples in detail. These examples are chosen in part to illustrate the variety of behaviors which are seen in SFAs. The calculations to support the conclusions are not presented here; the reader is encouraged to perform them.

The Golden Mean Process The first example is the *Golden Mean Process* (GMP) which we have already seen in section 2.6. This presentation is deterministic, and the recurrent process states and the presentation states coincide, so GMP is a stochastic deterministic finite automaton.

GMP has two symbols, and its smallest HMM presentation has two states. Its most prominent feature is that its output sequences never contain pairs of consecutive 0s. The reader should be able to verify that $\pi = (\frac{2}{3}, \frac{1}{3})$ from the transition matrices T^0 and T^1.

$$\begin{gathered} V = \{\mathbf{B}, \mathbf{C}\},\ \mathcal{X} = \{0, 1\},\ \pi = (\tfrac{2}{3}, \tfrac{1}{3}) \\ T^0 = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix},\ T^1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \end{pmatrix} \end{gathered} \tag{3.54}$$

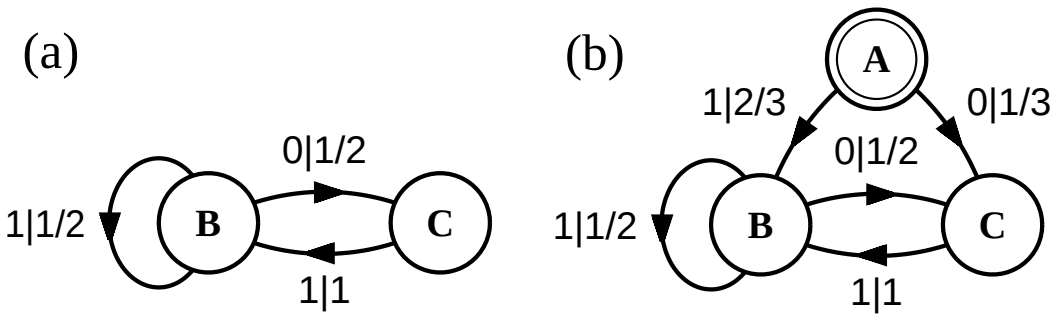

Fig. 3.3 (a) Labeled directed graph presentation of GMP. (b) Process state graph presentation of the process GMP. In this case, the recurrent process states coincide with the presentation states.

This process has three process states. If w is a history or history suffix which ends in 1, it induces process state $\mathbf{B}$. The mixed state it induces is $N(\pi T^w) = (1, 0)$. If w is a history or history suffix which ends in 0, it induces process state $\mathbf{C}$ and mixed state $(0, 1)$. This covers all histories and all history suffixes except λ, which induces process state $\mathbf{A}$, which is transient, and mixed state $\pi = (\frac{2}{3}, \frac{1}{3})$. The probabilities associated with the start state are $\mathbf{P}(1|\lambda) = \pi T^1 \vec{1} = \frac{2}{3}$ and $\mathbf{P}(0|\lambda) = \pi T^0 \vec{1} = \frac{1}{3}$. Similarly, the states to which these transitions are made are identified by comparing mixed states; $C_1(\pi) = (1, 0)$ and $C_0(\pi) = (0, 1)$. These are the mixed states associated to states $\mathbf{B}$ and $\mathbf{C}$, respectively.

The Simple Nondeterministic Source Our next example is the *Simple Nondeterministic Source (SNS)*, which we saw in section 3.2. This process can be represented with only two presentation states, but as we will see shortly, it has infinitely many process states. A two-state HMM presentation is

$$\begin{gathered} V = \{\mathbf{A}, \mathbf{B}\},\ \mathcal{X} = \{0, 1\},\ \pi = (\tfrac{1}{2}, \tfrac{1}{2}), \\ T^0 = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & 0 \end{pmatrix},\ T^1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} \end{pmatrix}. \end{gathered} \tag{3.55}$$

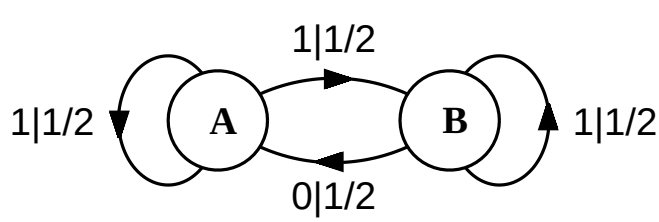

Fig. 3.4 Labeled directed graph presentation for SNS.

Let $w_n = 01^n$, the word consisting of a 0 followed by n 1s, and let $\mathbf{A}_n$ be the process state induced by w_n. The matrix corresponding to w_n is

$$T^{w_n} = T^0 \left(T^1\right)^n = \begin{pmatrix} 0 & 0 \\ 2^{-n-1} & n2^{-n-1} \end{pmatrix}, \tag{3.56}$$

so the mixed state $\mu(w_n)$ corresponding to $\mathbf{A}_n$ is

$$N(\pi T^{w_n}) = N\left(2^{-n-1}, n2^{-n-1}\right) = \left(\frac{1}{n+1}, \frac{n}{n+1}\right). \tag{3.57}$$

The first few of these states are listed in table 3.1. The $\mathbf{A}_n$ are all distinct states, since their mixed state versions are all distinct. In fact, the $\mathbf{A}_n$ comprise all but one of the process states. Also, the word 0 is a synchronizing word, since $\mathbf{P}(\cdot|0) = \mathbf{P}(\cdot|w0)$ for all words w such that $\mathbf{P}(w0) > 0$. We can verify this by calculating $C_0(\pi) = (1,0)$ and $C_0(\mu) = (1,0)$ for all μ. Thus all of the w_ns are synchronizing words, and all of the $\mathbf{A}_n$ are reachable recurrent states.

SNS is also an example of a process in which reachable recurrent states are induced by words which are not synchronizing. This precludes the possibility of a converse to proposition 2.6.2, which said that synchronizing words induce reachable recurrent states. The word 11 induces the process state $\mathbf{A}_3$, a reachable recurrent state, also induced by $w_3 = 0111$. However, 11 is not a synchronizing word, because $\mathbf{P}(\cdot|011)$ and $\mathbf{P}(\cdot|111)$ are not equal to $\mathbf{A}_3$.

History or history suffix s	Mixed state $\mu(s)$	Process state	$\mathbf{P}(\text{symbol } 0\|w)$
$\ldots 0$	$(1,0)$	$\mathbf{A}_0$	0
$\ldots 01$	$\left(\frac{1}{2}, \frac{1}{2}\right)$	$\mathbf{A}_1$	$\frac{1}{4}$
$\ldots 011$	$\left(\frac{1}{3}, \frac{2}{3}\right)$	$\mathbf{A}_2$	$\frac{1}{3}$
$\ldots 0111$	$\left(\frac{1}{4}, \frac{3}{4}\right)$	$\mathbf{A}_3$	$\frac{3}{8}$
$\ldots 01111$	$\left(\frac{1}{5}, \frac{4}{5}\right)$	$\mathbf{A}_4$	$\frac{2}{5}$
$\ldots 01^n$	$\left(\frac{1}{n+1}, \frac{n}{n+1}\right)$	$\mathbf{A}_n$	$\frac{1}{2}\frac{n}{n+1}$
infinitely many 1s	$(0,1)$	$\mathbf{A}_\infty$	$\frac{1}{2}$

Table 3.1 The first few mixed and process states of the “simple nondeterministic source”.

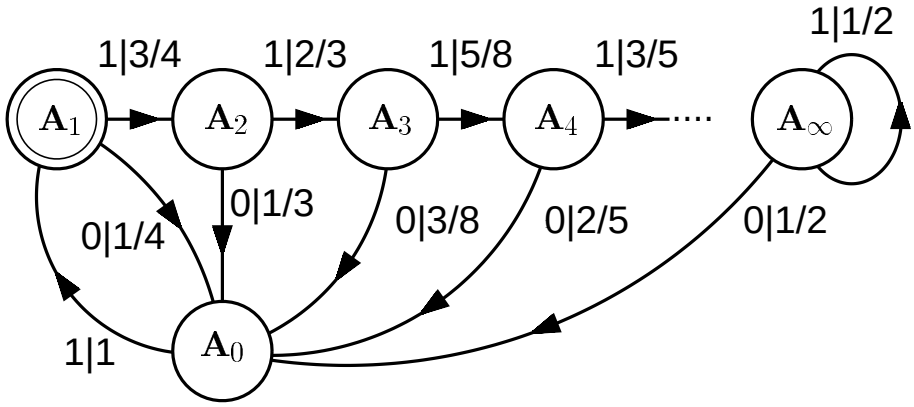

Fig. 3.5 An abbreviated version of the deterministic labeled directed graph presentation for the process "simple nondeterministic source," which has infinitely many process states.

The Cantor process Our third example is the *Cantor* process, which has the following HMM presentation:

$$V = \{\mathbf{A}, \mathbf{B}\},\ \mathcal{X} = \{0, 1\},\ \pi = (\tfrac{1}{2}, \tfrac{1}{2}),$$
$$T^0 = \begin{pmatrix} 0.55 & 0 \\ 0.30 & 0.15 \end{pmatrix},\ T^1 = \begin{pmatrix} 0.15 & 0.30 \\ 0 & 0.55 \end{pmatrix}. \tag{3.58}$$

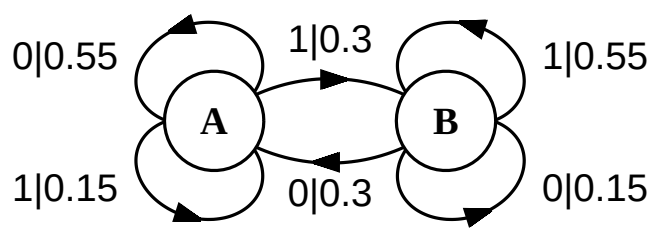

Fig. 3.6 Labeled directed graph presentation for the Cantor process.

Recall that a process state is an equivalence class of histories and history suffixes. For the Cantor process, all of these equivalence classes are trivial: every history and every history suffix induces a future conditional distribution which is different from that generated by every other history and every other history suffix. (Of course, pairs of future conditional distributions exist arbitrarily close to one another.) The result is that the Cantor process has uncountably many elusive process states, one induced by each history. Also, it has countably many strictly transient states, which are in one-to-one correspondence with the history suffixes. Thus, this is an example of a process with no synchronizing words.

Note that

$$N\big((x,1-x)T^1\big) = \left(\frac{3x}{11-2x}, \frac{11-5x}{11-2x}\right) \approx \left(\frac{x}{3}, 1-\frac{x}{3}\right), \text{ and}$$
$$N\big((1-y,y)T^0\big) = \left(\frac{11-5y}{11-2y}, \frac{3y}{11-2y}\right) \approx \left(1-\frac{y}{3}, \frac{y}{3}\right). \tag{3.59}$$

These approximations are exact at $(0,1)$ and $(1,0)$, and are within $\frac{1}{54}$ in between. Thus, if μ is the mixed state induced by a history s, appending a symbol to s corresponds approximately to moving μ two-thirds of the distance to either $(0,1)$ or $(1,0)$, respectively. The mixed states induced by histories form a set similar to the middle-thirds Cantor set, hence the process's name. This may be seen in figure 3.7, which is a plot of the Cantor process's mixed states, all of which lie on the line segment with endpoints $(0,1)$ and $(1,0)$. The mixed states induced by history suffixes lie in the middle of the intervals which are deleted to form the approximate Cantor set. (It is possible to construct an HMM for which the mixed states induced by histories are exactly the middle-thirds Cantor set, but it is degenerate — it is equivalent to a fair coin.)

Fig. 3.7 The mixed states for the process Cantor. Dots are mixed states corresponding to elusive process states. The small vertical lines are the mixed states corresponding to the subset of transient process states.

The Two Biased Coins process The last example we will look at here is the *Two Biased Coins (2BC)* process. The process can be simulated with a pair of biased coins. One of the biased coins is chosen by a flip of a fair coin. The chosen biased coin is then flipped to produce a bi-infinite sequence. Like the above examples, it has the following two-state presentation:

$$V = \{\mathbf{A}, \mathbf{B}\},\ \mathcal{X} = \{0,1\},\ \pi = (\tfrac{1}{2}, \tfrac{1}{2}),$$
$$T^0 = \begin{pmatrix} \frac{3}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix},\ T^1 = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{3}{4} \end{pmatrix}. \tag{3.60}$$

2BC is a reducible process, as it consists essentially of two processes with no interaction between them; see figure 3.8.

Calculation of 2BC's mixed states is equivalent to using Bayesian methods to infer which of the two biased coins is being flipped. The stationary distribution π is the prior

distribution, and $\mu(w)$ is simply the posterior distribution over the presentation states given the word w. The procedure we use for calculating $\mu(wx)$ from $\mu(w)$ can be viewed as a procedure to dynamically update the posteriors. If w is a word of length $i+j$ consisting of i 0s and j 1s in any order, then $T^w = \begin{pmatrix} \frac{3^i}{4^{i+j}} & 0 \\ 0 & \frac{3^j}{4^{i+j}} \end{pmatrix}$, and resulting mixed state $\mu(w) = N(\pi T^w)$ is $\frac{1}{3^i+3^j}\left(3^i, 3^j\right) = \frac{1}{3^{i-j}+1}\left(3^{i-j}, 1\right)$. So the mixed state — and the process state — induced by a word depends only on the difference between the number of 0s and the number of 1s in the word. Thus, there are countably infinitely many reachable process states, one for each integer. Some of these process state are portrayed in figure 3.8a.

With finite data, we are never sure which presentation state the process is in, so all reachable process states are transient. Asymptotically, as the length of the word goes to infinity, we can be sure with probability 1 which of the presentations states we are in. Thus, the mixed state versions of the recurrent process states are $(1,0)$ and $(0,1)$, Hence there are exactly two recurrent process states, both of which are unreachable and which correspond exactly to the presentation states.

There are also uncountably many histories in which the difference between the number of zeros and the number of ones is bounded, for example $\ldots 010101$. These histories induce elusive states, the total probability of which is 0.

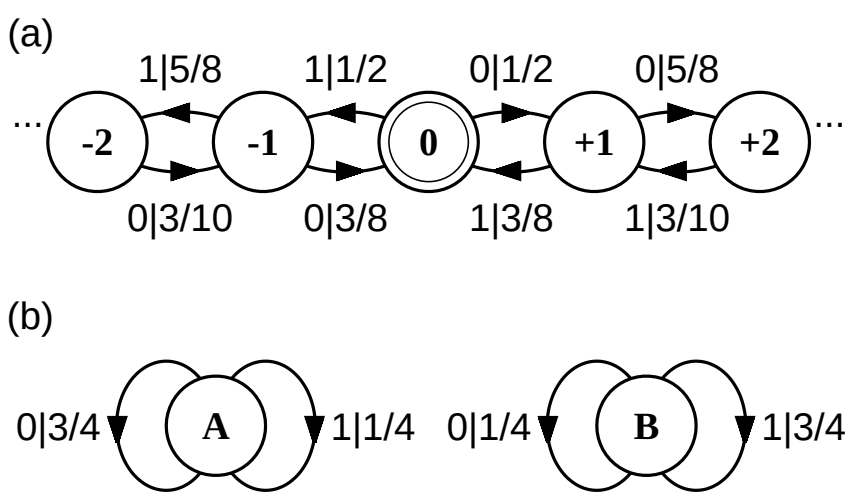

Fig. 3.8 (a) A subset of the transient process states for 2BC. The state induced by a word is determined solely by the number of 0s minus the number of 1s in the word. All of the transient states are infinitely preceded. (b) The recurrent states of 2BC. Both **A** and **B** are unreachable recurrent states. Each connected component of the graph corresponds to a recurrent component of the underlying Markov Chain.

3.6 When Does a Process have an HMM Presentation?

In this section we propose a characterization of when a process is an SFA — that is, when a process has a (finite) HMM presentation — and we show that it is a necessary condition. This characterization is a keystone of this dissertation. It leads almost directly to the reconstruction algorithm of chapter 5. And it follows from the following observation.

If we have a process with an HMM presentation $\left(V, \mathcal{X}, \{T^k\}, \pi\right)$, then a mixed state for that presentation is a distribution on V, or equivalently, a vector of $|V|$ components. This means that all the mixed states lie in the $|V|$-dimensional vector space $\mathbb{R}^{|V|}$. This in turn means that the dimension of the span of the mixed states is less than or equal to $|V| < \infty$. As process states are essentially equivalent to mixed states, we can make a similar statement about the process states.

First, we need to be able to work with process states as elements of a vector space. Let $\mathcal{W}$ be the set of all signed measures on the future space. These include the process states: if $\mathbf{A}$ is a process state, then $\mathbf{A} \in \mathcal{W}$. For any $\mathbf{A}, \mathbf{B} \in \mathcal{W}$ and $c, d \in \mathbb{R}$, we define $c\mathbf{A} + d\mathbf{B}$ as follows. For all future words w, $(c\mathbf{A} + d\mathbf{B})(w)$ is defined to be $c\mathbf{A}(w) + d\mathbf{B}(w)$, so that $\mathcal{W}$ is a vector space. In addition, if $\mathbf{A}, \mathbf{B}$ are probability measures and $c, d \geq 0,\ c + d = 1$, then $c\mathbf{A} + d\mathbf{B}$ is a probability measure.

We now state the main result of this section.

Theorem 3.6.1 Given a process $\mathcal{P}$, let $\mathcal{U}$ be the subspace of $\mathcal{W}$ spanned by the reachable process states. If the $\mathcal{P}$ has an HMM presentation $\left(V, \mathcal{X}, \{T^k\}, \pi\right)$, then dim$(\mathcal{U}) \leq |V|$.

Before we can readily prove this result, we need to develop the connection between $\mathcal{W}$ and $\mathbb{R}^{|V|}$.

Lemma 3.6.2 Suppose we have $\mathbf{A}, \mathbf{B} \in \mathcal{W}$ and $\mu, \nu \in \mathbb{R}^{|V|}$ such that for all future words w we have $\mu T^w \vec{1} = \mathbf{A}(w)$ and $\nu T^w \vec{1} = \mathbf{B}(w)$. Then for all $c, d \in \mathbb{R}$ and for all future words w, we have $(c\mu + d\nu)T^w \vec{1} = (c\mathbf{A} + d\mathbf{B})(w)$.

Proof. $(c\mu + d\nu)T^w \vec{1} = c\left(\mu T^w \vec{1}\right) + d\left(\nu T^w \vec{1}\right) = c\mathbf{A}(w) + d\mathbf{B}(w) = (c\mathbf{A} + d\mathbf{B})(w)$. ■

For the next lemma, we need some additional notation. We will use $\underline{0}$ to denote the zero vector in $\mathbb{R}^{|V|}$ (Note that $\underline{0}$ is a row vector, in contrast to $\vec{1}$.) Also, we will use $\mathbf{0}$ to

denote the zero measure in $\mathcal{W}$. Thus we have, for all future words w, $\underline{0}T^w\vec{1} = \mathbf{0}(w) = 0$.

Lemma 3.6.3 Suppose we have reachable process states $\mathbf{A}_1, \ldots, \mathbf{A}_l \in \mathcal{W}$, and vectors $\mu^1, \ldots, \mu^l \in \mathbb{R}^{|V|}$ such that μ^i is a mixed state version of $\mathbf{A}_i$ for each $i \in 1, \ldots, l$. If there exist real numbers $c_1, \ldots, c_l$, not all zero, such that $\sum_{i=1}^{l} c_i\mu^i = \underline{0}$, then $\sum_{i=1}^{l} c_i\mathbf{A}_i = \mathbf{0}$.

Proof. For all future words w, we have

$$\left(\sum_{i=1}^{k} c_i\mathbf{A}_i\right)(w) = \left(\sum_{i=1}^{k} c_i\mu^i\right)T^w\vec{1} \tag{3.61}$$

by lemma 3.6.2. However, the right hand side of equation 3.61 is zero by assumption. Thus the left hand side is also zero for all w, and we have

$$\sum_{i=1}^{k} c_i\mathbf{A}_i = \mathbf{0}. \blacksquare \tag{3.62}$$

Proof of theorem 3.6.1 Choose any $|V| + 1$ reachable process states $\mathbf{A}_1, \ldots, \mathbf{A}_{|V|+1}$, and choose $\mu_1, \ldots, \mu_{|V|+1} \in \mathbb{R}^{|V|}$ such that μ_i is a mixed state version of $\mathbf{A}_i$ for each $i \in 1, \ldots, k$. The μ_i are a set of $|V| + 1$ vectors in a $|V|$-dimensional vector space, so they must be dependent. That is, there must exist $c_1, \ldots, c_{|V|+1}$ such that $\sum_{i=1}^{|V|+1} c_i\mu_i = \underline{0}$. Now, by lemma 3.6.3, $\sum_{i=1}^{|V|+1} c_i\mathbf{A}_i = 0_{\mathcal{W}}$, so the $\mathbf{A}_i$s are linearly dependent. Thus, we have shown that a set of linearly independent process states has size at most $|V|$, so the span of the process states is at most $|V|$-dimensional. $\blacksquare$

The following fact about SFAs follows immediately from theorem 3.6.1.

Corollary 3.6.4. Given a process, let $\mathcal{U}$ be the span of its process states. If it has a (finite) HMM presentation, then dim$(\mathcal{U}) < \infty$.

Proof. For any finite HMM $\left(V, \mathcal{X}, \{T^k\}, \pi\right)$, we have $|V| < \infty$. Thus, using theorem 3.6.1, we have dim$(\mathcal{U}) \leq |V| < \infty$.$\blacksquare$

As an illustration of the use of corollary 3.6.4, we will now prove the following statement, which was stated without proof at the end of section 3.3.

Proposition 3.6.5. The modified nested parentheses process does not have an HMM presentation.

Proof. Let $w_n = !(^{n-1}$ be the length n word consisting of a ! followed by $n-1$ (s. Similarly, let $s_n =)^{n-1}!$ be the mirror image of w_n. After the word w_n, the counter must be $n-1$. Before the word s_m, the counter must be $m-1$. Thus s_m cannot follow w_n if $m \neq n$:

$$\mathbf{P}(s_m|w_n) = \begin{cases} \left(\frac{2}{3}\right)^n & m = n \\ 0 & m \neq n. \end{cases} \tag{3.63}$$

Now let $\mathbf{A}_n$ be the process state induced by w_n,

$$\mathbf{A}_n(s_m) = \begin{cases} \left(\frac{2}{3}\right)^n & m = n \\ 0 & m \neq n. \end{cases} \tag{3.64}$$

For every n and any for linear combination $c_1, \ldots, c_{n-1}$, we have $(c_1\mathbf{A}_1 + \ldots c_{n-1}\mathbf{A}_{n-1})(s_n) = 0$, while $\mathbf{A}_n(s_n) > 0$. Thus, $\mathbf{A}_n$ is linearly independent of $\mathbf{A}_1, \ldots, \mathbf{A}_{n-1}$. In this way we see that we can construct arbitrarily large, linearly independent sets of process states. Thus we have $\dim(\text{span}\{\mathbf{A}_i | i > 0\}) = \infty$, so this process cannot have an HMM presentation. ■

A related condition for functions of Markov chains was shown by Gilbert [20], and variants appear in [10] and [21]. These conditions are stated in terms of a different context of definitions and terminology, so that their exact relationship to corollary 3.6.4 is difficult to ascertain. The author suspects that if one developed the appropriate machinery to connect these contexts, one would find that the conditions are equivalent.

We have shown that $\dim(\mathcal{U}) < \infty$ is a necessary condition for a process to have an HMM presentation. It is almost a sufficient condition. In order to make this precise, however, we will need to develop a generalization of HMMs. This generalization is the subject of the next chapter.