4 Generalized Hidden Markov Models

At the beginning of the last chapter, we viewed a Hidden Markov Model as a Markov chain with a stochastic output filter, and we only interpreted the presentation states as the states of a Markov Chain. Over the course of that chapter, we used the tools of linear algebra more and more, working with mixed states rather than directly with presentation states. Finally, we introduced an alternative interpretation of the presentation states — that they are basis elements for the set of mixed states, or equivalently, for the set of process states. In this interpretation, the transformation matrices define the process by defining linear transformations on the mixed states. As we will see, this linear algebra interpretation is the more fundamental one.

In this chapter, we will define and use a generalization of Hidden Markov Models in which the presentation states cannot be interpreted as states of a Markov chain, but can only reasonably be interpreted as basis elements. The first use of Generalized Hidden Markov appears to have been as a counter-example in [10]. Several authors have used Generalized Hidden Markovs and similar techniques in recent years, including connection to neural nets developed in [22] and the solution of the problem of HMM equivalence in [7]. Late in this chapter, and in the next one, we will choose a linearly independent basis for the span of the process states, and use these basis vectors as presentation states of a new presentation we construct directly from the process.

4.1 Generalized Hidden Markov Models

When we think of a Hidden Markov Model as an object of linear algebra, it makes sense to consider what happens when we perform a change of basis — a canonical linear algebra operation. And so, after one convenient definition, we will work through a change of basis for a generic HMM.

Definition 4.1.1. A *unit-sum* vector is a vector whose components sum to one. A unit-sum matrix is a matrix whose row vectors are unit-sum vectors.

That is, a row vector v is unit-sum row vector if $v\vec{1} = 1$, and a matrix A is a unit sum matrix if

$$A\vec{1} = \vec{1}. \tag{4.1}$$

Note that the inverse of a unit-sum matrix must also be unit-sum. If we multiply both sides of equation 4.1 by A^{-1}, we get

$$I\vec{1} = A^{-1}\vec{1} \tag{4.2}$$

and, clearly, $I\vec{1} = \vec{1}$. If a unit-sum vector satisfies the additional requirement that all of its components are nonnegative, then it is a *stochastic* vector. Similarly a matrix is stochastic if all of its rows are stochastic vectors.

Suppose we have an HMM $\mathcal{I} = \left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$. A mixed state $\mu = \left(\mu_1, \ldots, \mu_{|V|}\right)$ for this HMM is a stochastic vector in a vector space with basis elements associated to the states $i \in V$: if μ is induced by a history object s, then $\mu_i = \mathbf{P}(i|s)$. The process state associated with μ is a linear combination $\sum_i \mu_i \mathbf{A}_i$ of the conditional future distributions $\mathbf{A}_i = \mathbf{P}(\cdot|i)$. If we let $\mathcal{U}$ be the span of all of the HMM's reachable process states, then $\left\{\mathbf{A}_1, \ldots, \mathbf{A}_{|V|}\right\}$ is the basis we used for $\mathcal{U}$ in section 3.6.

We will now work through a change of basis. We begin by choosing a new basis $\left\{\mathbf{B}_1, \ldots, \mathbf{B}_{|V|}\right\}$ for $\mathcal{U}$ as follows: choose an invertible unit-sum $|V| \times |V|$ matrix M, and for all i let

$$\mathbf{B}_i = \sum_j M_{ij} \mathbf{A}_j. \tag{4.3}$$

Clearly, for all i, $\mathbf{B}_i \in \mathcal{U}$, and because M is invertible, $\left\{\mathbf{B}_1, \ldots, \mathbf{B}_{|V|}\right\}$ is a basis for $\mathcal{U}$.

This change of basis calculation will be facilitated by the following somewhat nonstandard notation. We will write the vectors of the basis in a formal column vector *as if they were scalars*. That is, we define the formal column vectors $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ by

$$\vec{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_{|V|} \end{pmatrix} \text{ and } \vec{\mathbf{B}} = \begin{pmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_{|V|} \end{pmatrix} \tag{4.4}$$

Thus, we may rewrite 4.3 as

$$\vec{\mathbf{B}} = M\vec{\mathbf{A}}. \tag{4.5}$$

In this notation, if $\mu \in \mathbb{R}^{|V|}$ is a mixed state, then it is a row vector, and it describes the process state $\mu\vec{\mathbf{A}}$. From equation 4.5, we get $M^{-1}\vec{\mathbf{B}} = \vec{\mathbf{A}}$, so $\mu M^{-1}\vec{\mathbf{B}} = \mu\vec{\mathbf{A}}$. Thus if $\nu = \mu M^{-1}$, then $\mu\vec{\mathbf{A}} = \nu\vec{\mathbf{B}}$, so μ and ν describe the same process state in different

coordinate systems. Hence M^{-1} maps $\vec{\mathbf{A}}$–coordinates to $\vec{\mathbf{B}}$–coordinates. Since ν is not a mixed state for any HMM we have defined yet, we will call it a coordinate vector.

Our HMM's stationary distribution π and its transition matrices are all given in $\vec{\mathbf{A}}$–coordinates. What do they look like in $\vec{\mathbf{B}}$–coordinates? π is simply a mixed state, so it transforms to the coordinate vector $\tau = \pi M^{-1}$ as we have indicated above. If ν is a coordinate vector, then $\nu\vec{\mathbf{B}} \in \mathcal{U}$, and νM defines a row vector in $\vec{\mathbf{A}}$ coordinates — a mixed state. We can operate on νM with the operator T^k, and transform the result back to $\vec{\mathbf{B}}$ coordinates with M^{-1}. The result is that the operation $\mu \to \mu T^k$ in $\vec{\mathbf{A}}$ coordinates becomes $\nu \to \nu M T^k M^{-1}$ in $\vec{\mathbf{B}}$ coordinates. That is, the similarity transformation which transforms T^k into $\vec{\mathbf{B}}$ coordinates produces the matrix $U^k = MT^kM^{-1}$.

Now, if we define a set of formal symbols $V' = \{1', 2', \ldots, |V|'\}$, we can construct a quadruple $(V', \mathcal{X}, \{U^k\}, \tau)$. This quadruple looks like an HMM. It may fail to be one, however, because the U^k matrices may have negative entries. Nonetheless, it satisfies the rest of the definition of an HMM. The transition matrices $\{U^k\}$ satisfy $\left(\sum_k U^k\right)\vec{1} = \vec{1}$, since

$$\begin{aligned}\left(\sum_k U^k\right)\vec{1} &= \left(\sum_k MT^kM^{-1}\right)\vec{1} \\ &= M\left(\sum_k T^k\right)M^{-1}\vec{1} \\ &= M\left(\sum_k T^k\right)\vec{1} \\ &= M\vec{1} = \vec{1}.\end{aligned} \tag{4.6}$$

And τ satisfies $\tau = \tau\sum_k U^k$ since

$$\begin{aligned}\tau\sum_k U^k &= \pi M^{-1}\sum_k MT^kM^{-1} \\ &= \pi M^{-1}M\left(\sum_k T^k\right)M^{-1} \\ &= \pi\left(\sum_k T^k\right)M^{-1} \\ &= \pi M^{-1} = \tau.\end{aligned} \tag{4.7}$$

Further, if we manipulate $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ as if it were an HMM, and calculate $\tau U^w \vec{1}$ for an arbitrary word $w = w_1 \ldots w_l$, we get

$$\begin{aligned} \tau U^w \vec{1} &= \tau U^{w_1} \ldots U^{w_l} \vec{1} \\ &= \left(\pi M^{-1}\right)\left(M U^{w_1} M^{-1}\right) \ldots \left(M U^{w_1} M^{-1}\right) \vec{1} \\ &= \pi T^{w_1} \ldots T^{w_l} M^{-1} \vec{1}, \end{aligned} \tag{4.8}$$

and since M^{-1} is unit-sum, this becomes

$$\tau U^w \vec{1} = \pi T^w \vec{1}. \tag{4.9}$$

Thus, for every word $w \in \mathcal{X}^*$, we get $\mathbf{P}(w) = \tau U^w \vec{1}$. In spite of the fact that it is not an HMM, $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ defines a process as if it were. We will call it a *Generalized Hidden Markov Model (GHMM)*, following [8].

For instance, consider the following presentation of the Golden Mean Process which we saw in section 3.5:

$$\begin{gathered} V = \{\mathbf{B}, \mathbf{C}\},\ \mathcal{X} = \{0, 1\},\ \pi = \left(\tfrac{2}{3}, \tfrac{1}{3}\right) \\ T^0 = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix},\ T^1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \end{pmatrix} \end{gathered} \tag{4.10}$$

As discussed in section 3.5, the processes states for this HMM are represented by the mixed states $\left(\frac{2}{3}, \frac{1}{3}\right)$, $(1, 0)$ and $(0, 1)$. Let

$$M = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}, \tag{4.11}$$

which is an invertible, unit-sum matrix. Note that $M^{-1} = M$. Now, when we perform the change of basis, we get

$$\tau = \pi M^{-1} = \left(\tfrac{4}{3}, -\tfrac{1}{3}\right), \tag{4.12}$$

$$U^0 = M T^0 M^{-1} = \begin{pmatrix} 1 & -\frac{1}{2} \\ 2 & -1 \end{pmatrix}, \text{ and} \tag{4.13}$$

$$U^1 = M T^2 M^{-1} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix}. \tag{4.14}$$

The new coordinates for the process states, which are the images of the mixed states under multiplication (on the right) by M^{-1}, are $\left(\frac{4}{3}, -\frac{1}{3}\right)$, $(1, 0)$, and $(2, -1)$.

How can we make sense of $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$? If we try to think of $1' \in V'$ as a state in some variation on a Markov chain, it makes no sense at all — this casts $\sum_k U^k$

in the role of a transition matrix, so we find ourselves looking at negative transition probabilities. If we insist on this interpretation, then we must reject the whole idea of GHMMs as absurd. But if we think of $1'$ as an vector in a basis for $\mathcal{U}$, there is no substantial difficulty. A row of U^k simply gives the coordinates of the image of some basis element under a linear mapping, and a negative coordinate is a perfectly sensible thing. Instead of contemplating possible meanings for negative probabilities, we simply stop interpreting the matrix entries U^k_{ij} as probabilities. Essentially, we attribute meaning to the entire matrix U^k — it is a linear map — but not to individual entries in this matrix. (We will continue to use the term *mixed state*, although it is no longer apt.)

Definition 4.1.2. A *Proto-Generalized Hidden Markov Model (Proto-GHMM)* is a quadruple $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$, where V and $\mathcal{X}$ are finite sets, and each T^k is a $|V| \times |V|$ matrix, and the following conditions are satisfied:

1. $\sum_k T^k$ is a unit-sum matrix*,
2. π is a unit-sum vector, and
3. $\pi = \pi \sum_k T^k$.

We would like to have a Proto-GHMM define a process in the same way that an HMM does, but this does not always happen, because of a complication introduced by allowing negative entries in the T^ks. Proto-GHMMs $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and words w exist such that $\pi T^w \vec{1} < 0$. An example is

$$\begin{gathered} V = \left\{0', 1'\right\},\ \mathcal{X} = \{0, 1\},\ \pi = \left(\frac{1}{2}, \frac{1}{2}\right) \\ T^0 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},\ T^1 = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}, \end{gathered} \tag{4.15}$$

for which $\pi T^0 \vec{1} = -1$. Clearly, then, a Proto-GHMM may fail to define a process. The next two definitions address this problem.

Definition 4.1.3. A Proto-GHMM $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ is *valid* if, for all words $w \in \mathcal{X}^*$, it satisfies $\pi T^w \vec{1} \geq 0$.

---

*The reader may wonder why we require $\sum_k U^k$ to be unit-sum, when we could discard this restriction and have greater generality. This extra generality costs us some convenience. For example, the fact that π and $\sum_k T^k$ are unit-sum guarantees that $\sum_k \mathbf{P}(k) = 1$. We can work around such difficulties, but there is no point — as we will see in chapter 5. Every process which could possibly be represented with this greater generality has a GHMM presentation.

Definition 4.1.4. A *Generalized Hidden Markov Model (GHMM)* is a valid Proto-GHMM.

This definition is precisely what we need in order to have GHMMs represent processes.

Proposition 4.1.5. Every GHMM defines a process.

Proof. We prove this by applying B.1.1, where $f(w) = \pi T^w \vec{1}$. Thus we must verify

1. $f(\lambda) = 1$, and
2. for all words $w \in \mathcal{X}^*$,

$$f(w) = \sum_{z \in \mathcal{X}} f(zw) = \sum_{z \in \mathcal{X}} f(wz). \tag{4.16}$$

Unlike previous applications of B.1.1, here we must also show that $f : \mathcal{X}^* \to [0, 1]$.

First, $f(\lambda) = \pi T^\lambda \vec{1}$, where T^λ is the identity matrix by definition and π is unit-sum. So clearly, $f(\lambda) = 1$. Next, we deal with $f(wz)$:

$$\begin{aligned} \sum_{z \in \mathcal{X}} f(wz) &= \sum_{z \in \mathcal{X}} \pi T^w T^z \vec{1} \\ &= \pi T^w \left(\sum_{z \in \mathcal{X}} T^z \right) \vec{1}. \end{aligned} \tag{4.17}$$

But $\sum_{z \in \mathcal{X}} T^z$ is a unit-sum matrix, so this becomes

$$\sum_{z \in \mathcal{X}} f(wz) = \pi T^w \vec{1} = f(w). \tag{4.18}$$

The other equality in equation 4.16 may be handled by a similar calculation, using $\pi = \pi \sum_{z \in \mathcal{X}} T^z$ in place of the unit-sum property.

Finally, we need to show that for an arbitrary word $w \in \mathcal{X}^*$, $0 \leq f(w) \leq 1$. Half of this is given by validity. Given the properties we have just shown, a simple induction argument establishes that for all l,

$$\sum_{s \in \mathcal{X}^l} f(s) = 1. \tag{4.19}$$

Let $l = |w|$, and rewrite 4.19 as

$$1 = \sum_{s \in \mathcal{X}^l} f(s) = f(w) + \sum_{s \in \mathcal{X}^l,\, s \neq w} f(s). \tag{4.20}$$

Both terms in this sum are nonnegative, so neither can be greater than 1.■

We can characterize the preceding development as follows: A Proto-GHMM is an object which would be an HMM if it didn't have negative entries in its matrices, and a GHMM is a Proto-GHMM which never assigns a negative number to a word, and thus defines a process. That is, we are allowing negative entries in transition matrices, but only when the result works with the procedures we use for HMMs.

Testing the validity of a Proto-GHMM is nontrivial. Consider the obvious, naive algorithm: Take the (countably infinite) list of all words in $\mathcal{X}^*$, and write a loop which computes $\pi T^w \vec{1}$ for every word w on the list. Make the loop halt if this quantity is negative, and continue down the list if it is not. The Proto-GHMM is valid if and only if the loop never halts. Clearly this is not a practical test. One can find improvements to this algorithm which prune this list and thus typically reach invalid conclusions faster. And there are some special cases in which validity can be established — for instance, every HMM is a (valid) GHMM. Nevertheless, the essence of the test in the general case remains the same.

We have now finished defining GHMMs, and we will present some results involving them. The first of these is an extension of theorem 3.6.1 to GHMMs. The proof of 3.6.1 will serve as a proof of 4.1.6 without modification, so we will not give a separate proof here. Recall that $\mathcal{W}$ is the set of all signed measures on the future.

Proposition 4.1.6. Given a process $\mathcal{P}$, let $\mathcal{U}$ be the subset of $\mathcal{W}$ spanned by the reachable process states. If $\mathcal{P}$ has a GHMM presentation $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$, then $\dim(\mathcal{U}) \leq |V|$.

We conclude this section by restating and expanding on the basis change manipulation we performed earlier in this section. We begin with the following result.

Proposition 4.1.7. If $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ is a GHMM and M is an invertible unit-sum matrix, and V' is any set of size $|V|$, then $\left(V', \mathcal{X}, \left\{MT^kM^{-1}\right\}, \pi M^{-1}\right)$ is a GHMM which defines the same process as $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$.

Proof. It may easily be verified that $\left(V', \mathcal{X}, \left\{MT^kM^{-1}\right\}, \pi M^{-1}\right)$ is a Proto-GHMM. And by the same arguments used above for conjugation of HMMs, we know that for any $w \in \mathcal{X}^*$,

$$\tau U^w \vec{1} = \pi T^w \vec{1}. \tag{4.21}$$

The right side of equation 4.21 is always nonnegative, so the left side must be also. Therefore $\left(V', \mathcal{X}, \left\{MT^kM^{-1}\right\}, \pi M^{-1}\right)$ is valid, and thus it is also a GHMM.■

Definition 4.1.8. We say that two Proto-GHMMs $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ are *conjugate* to each other by an invertible unit-sum matrix M if

1. $|V| = |V'|$,
2. $\tau = \pi M^{-1}$, and
3. for all k, $U^k = MT^kM^{-1}$.

Note that this is a linear conjugacy, which is the only kind of conjugacy we will consider.

Proposition 4.1.7 tells us, then, that if a Proto-GHMM is conjugate to a GHMM, then it is itself a GHMM.

Definition 4.1.9. When two GHMMs define the same process, we say that they are *equivalent*.

Thus the proof of proposition 4.1.7 also shows that if two GHMMs are conjugate, then they are equivalent.

Lemma 4.1.10. If two GHMMs $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$, are conjugate by an invertible unit-sum matrix M, then for all $w \in \mathcal{X}^*$,

$$\tau U^w M = \pi T^w. \tag{4.22}$$

Note that since these two HMMs are equivalent, we know that $\pi T^w \vec{1} = \tau U^w \vec{1}$. So if we divide equation 4.22 by this quantity, we get

$$\frac{\tau U^w}{\tau U^w \vec{1}} M = \frac{\pi T^w}{\pi T^w \vec{1}}, \tag{4.23}$$

which we may recognize as

$$N(\tau U^w) M = N(\pi T^w). \tag{4.24}$$

That is, M takes mixed states to mixed states.

Proof. We are given that $U^0 = MT^0M^{-1}$, $U^1 = MT^1M^{-1}$, and $\tau = \pi M^{-1}$. Multiply by M, and we have $MT^0 = U^0M$, $MT^1 = U^1M$, and $\pi = \tau M$. Thus, for all w,

$$\begin{aligned} \tau U^w M &= \tau U^{w_1} \dots U^{w_l} M \\ &= \tau U^{w_1} \dots U^{w_{l-1}} M T^{w_l} \\ &\vdots \\ &= \tau M T^{w_1} \dots T^{w_l} \\ &= \pi T^w. \quad \blacksquare \end{aligned} \tag{4.25}$$

The converse of proposition 4.1.7 does not hold — there are pairs of GHMMs which are equivalent but not conjugate. This is caused by redundancy — extra states in the presentation. In fact, there are pairs of such presentations which are both HMMs. Consider

$$\begin{aligned} V = \{0,1,2\},\ \mathcal{X} = \{0,1\},\ \pi = \left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right) \\ T^0 = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix},\ T^1 = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \end{pmatrix}, \end{aligned} \tag{4.26}$$

and

$$\begin{aligned} V' = \{0', 1', 2'\},\ \mathcal{X} = \{0,1\},\ \pi = \left(\tfrac{2}{9}, \tfrac{1}{3}, \tfrac{4}{9}\right) \\ U^0 = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix},\ U^1 = \begin{pmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \end{pmatrix}. \end{aligned} \tag{4.27}$$

These are both presentations of the Golden Mean Process, which are redundant in different ways. The reader may see this by noticing that states 0 and 2 — and $0'$ and $2'$ — have the same future conditional distributions. Thus, 0 and 2 — and $0'$ and $2'$ —can be merged.

To show that these presentations are not conjugate, we will show that there exists no invertible unit-sum matrix A which takes πT^w to τU^w for all $w \in \mathcal{X}^*$. Lemma

w	πT^w	τU^w
0	$\left(0, \frac{1}{3}, 0\right)$	$\left(0, \frac{1}{3}, 0\right)$
01	$\left(0, 0, \frac{1}{3}\right)$	$\left(0, 0, \frac{1}{3}\right)$
011	$\left(\frac{1}{6}, 0, 0\right)$	$\left(\frac{1}{6}, 0, 0\right)$
0111	$\left(\frac{1}{12}, 0, 0\right)$	$\left(0, 0, \frac{1}{12}\right)$

Table 4.1 The actions of the HMMs given in equations 4.26 and 4.27 on selected words.

4.1.10 tells us that each row of table 4.1 constrains the matrix A. These constraints are incompatible; the first three rows imply that A is the identity matrix, and the fourth implies that it is not. Thus, no suitable matrix A exists, so these presentations are not conjugate.

4.2 Redundancy and Linear Algebra

In the last section we saw an example in which two equivalent presentations may fail to be conjugate to one another, because they are redundant in different ways. We will now study redundancy, and then return in the next section to GHMM equivalence. The methods we will develop here give us a new way of describing the essential information in a GHMM. In this new form, we will be able to identify and factor out redundancy, which is the key to resolving the equivalence and minimization problems.

Vector Spaces We begin by identifying two vector spaces $\mathcal{H}$ and $\mathcal{F}$, which we will call the history and future spaces. Given a GHMM $\left(V, \mathcal{X}, \{T^k\}, \pi\right)$, let $\mathcal{H}$ be the span of the set of all mixed states:

$$\mathcal{H} = \text{span}\{N(\pi T^w) | w \in \mathcal{X}^*\}. \tag{4.28}$$

In a complementary fashion, let

$$\mathcal{F} = \text{span}\left\{T^s \vec{1} | s \in \mathcal{X}^*\right\}. \tag{4.29}$$

Elements of $\mathcal{H}$ are linear combinations of mixed states, which are row vectors, and elements of $\mathcal{F}$ are column vectors. Just as $N(\pi T^w)$ contains all the information about the history suffix w that is relevant to the future, $T^s \vec{1}$ contains all the information about the future that is relevant to the past. Implicit in this is that π and $\vec{1}$ play analogous roles which is suggested by the identities $\pi \sum_k T^k = \pi$ and $\sum_k T^k \vec{1} = \vec{1}$. Just as we may use πT^w to calculate the conditional distributions on the future induced by w, we may use $T^s \vec{1}$ to calculate the conditional distributions on the past induced by s. Thus, we may think of $T^s \vec{1}$ as a backward analog of a process state, and we may think of any $f \in \mathcal{F}$ as a linear combination of these.

Let $h = N(\pi T^w)$ and $f = T^s \vec{1}$. If we take the product of these, we get $hf = \mathbf{P}(s|w)$. In considering the product hf, we may think of h as the linear functional and f as the

operand, or vice versa. In general, if $h \in \mathcal{H}$ and $f \in \mathcal{F}$, then we may think of h as a map $h : \mathcal{F} \to \mathbb{R}$ defined by $h(f) = hf$, and we may think of f as a map $f : \mathcal{H} \to \mathbb{R}$ given by $f(h) = hf$. Thus $\mathcal{H}$ and $\mathcal{F}$ are almost each other's dual spaces.

But $\mathcal{H}$ and $\mathcal{F}$ may not be each other's duals. If there is redundancy in the presentation states — that is, if there are distinct mixed states which induce the same conditional future — then it may happen that $\mathcal{H}$ and $\mathcal{F}$ have different dimensions, and there may be nonzero $h \in \mathcal{H}$ for which $hf = 0$ for all $f \in \mathcal{F}$.

Let

$$K_{\mathcal{F}} = \{h \in \mathcal{H} | \text{ for all } f \in \mathcal{F},\ hf = 0\}, \tag{4.30}$$

and similarly,

$$K_{\mathcal{H}} = \{f \in \mathcal{F} | \text{ for all } h \in \mathcal{H},\ hf = 0\}. \tag{4.31}$$

That is, $h \in \mathcal{H}$ is in $K_{\mathcal{F}}$ if it is in the kernel of every $f \in \mathcal{F}$. If $h = \pi T^w$ is a row vector induced by a word w and $h \in K_{\mathcal{F}}$, we usually have $h = (0, \ldots, 0)$ and $\mathbf{P}(w) = 0$. In an HMM, this is the only way a row vector induced by a word may fall in $K_{\mathcal{F}}$. But differences between mixed states may lie in $K_{\mathcal{F}}$. Suppose two words w_1 and w_2 induce the same process state, $\mathbf{P}(\cdot|w_1) = \mathbf{P}(\cdot|w_2)$, and let $h_1 = N(\pi T^{w_1})$ and $h_2 = N(\pi T^{w_2})$. Then for all words s,

$$h_1 T^s \vec{1} = \mathbf{P}(s|w_1) = \mathbf{P}(s|w_2) = h_2 T^s \vec{1}. \tag{4.32}$$

Because every $f \in \mathcal{F}$ is a linear combination of column vectors $T^s\vec{1}$, this implies $h_1 f = h_2 f$ for all $f \in \mathcal{F}$, or $h_1 - h_2 \in K_{\mathcal{F}}$. If we know that the current history suffix is either w_1 or w_2, but we do not know which one, then our finding out which one does not improve our ability to predict the future. And this works backward, too — if $h_1 - h_2 \in K_{\mathcal{F}}$, then w_1 and w_2 induce the same process state. $K_{\mathcal{F}}$ contains exactly those vectors which are irrelevant to the future of the process. For this reason we say that $K_{\mathcal{F}}$ consists of redundancy.

Similar statements are true about $K_{\mathcal{H}}$. If, and only if, $f_1, f_2 \in \mathcal{F}$ and $hf_1 = hf_2$ for all $h \in \mathcal{H}$, then $f_1 - f_2 \in K_{\mathcal{H}}$, and the distinction between f_1 and f_2 is independent of the history of the process.

Now we can eliminate this redundancy — factor it out, so to speak — by working with the quotient spaces $\mathcal{H}/K_{\mathcal{F}}$ and $\mathcal{F}/K_{\mathcal{H}}$ in place of $\mathcal{H}$ and $\mathcal{F}$. As the following

lemma shows, $\mathcal{H}/K_{\mathcal{F}}$ is in one-to-one correspondence with $\mathcal{U}$ the span of the reachable process states, so it contains no redundancy.

Lemma 4.2.1. *$\mathcal{H}/K_{\mathcal{F}}$ is isomorphic to $\mathcal{U}$.*

Proof. Let ψ be the quotient map $\psi : \mathcal{H} \to \mathcal{H}/K_{\mathcal{F}}$, and let $M : \mathcal{H} \to \mathcal{U}$ be defined as follows: if $h \in \mathcal{H}$, then $M(h)$ is the signed measure on the future space given by $M(h)(s) = hT^s\vec{1}$ for all $s \in \mathcal{X}^*$. Now we can define our isomorphism $\phi : \mathcal{H}/\mathcal{K}_{\mathcal{F}} \to \mathcal{U}$. If $g \in \mathcal{H}/K_{\mathcal{F}}$, choose $h \in \psi^{-1}(g)$ and define $\phi(g) = M(h)$. The value of $\phi(g)$ does not depend on our choice of h, because if $h_1, h_2 \in \psi^{-1}(g)$, then $h_1 - h_2 \in K_{\mathcal{F}}$. This implies that $h_1T^s\vec{1} = h_2T^s\vec{1}$ for all s, so $M(h_1) = M(h_2)$.

Because ψ and M are linear, ϕ must be linear. To show that it is an isomorphism, we must show that it is injective and surjective. The first of these is trivial: suppose $\phi(g_1) = \phi(g_2)$. If we choose $h_1 \in \psi^{-1}(g_1)$ and $h_2 \in \psi^{-1}(g_2)$, we have $M(h_1) = M(h_2)$, which is equivalent to $h_1T^s\vec{1} = h_2T^s\vec{1}$ for all s. Thus $h_1 - h_2 \in K_{\mathcal{F}}$, so $g_1 = g_2$.

Next we show that ϕ is surjective. The definition of $\mathcal{U}$ implies that there must exist words $w_1, \ldots, w_n$ such that the process states $\mathbf{A}_1 = \mathbf{P}(\cdot|w_1), \ldots, \mathbf{A}_n = \mathbf{P}(\cdot|w_n)$ form a linearly independent basis for $\mathcal{U}$. That is, there are no real numbers $a_1, \ldots, a_n$, such that

$$(a_1\mathbf{A}_1 + \ldots + a_n\mathbf{A}_n)(s) = 0 \text{ for all } s. \tag{4.33}$$

For $i = 1, \ldots, n$, let $g_i = \psi(N(\pi T^{w_i}))$. For all i and for all s, we have

$$\begin{aligned} \phi(g_i)(s) &= M(N(\pi T^{w_i}))(s) \\ &= \mathbf{P}(s|w) \\ &= \mathbf{A}(s), \end{aligned} \tag{4.34}$$

so $\phi(g_i) = \mathbf{A}_i$. Thus if $g_1, \ldots, g_n$ were linearly dependent, $\mathbf{A}_1, \ldots, \mathbf{A}_n$ would have to be linearly dependent as well. So $g_1, \ldots, g_n$ are a linearly independent basis for some subspace of $\mathcal{H}/K_{\mathcal{F}}$ and ϕ is a map which takes this basis to a basis for $\mathcal{U}$. This means that ϕ is an isomorphism from this subspace onto $\mathcal{U}$. But ϕ is injective and takes all of $\mathcal{H}/K_{\mathcal{F}}$ to $\mathcal{U}$, hence this subspace is all of $\mathcal{H}/K_{\mathcal{F}}$.■

An element of $\mathcal{H}/K_{\mathcal{F}}$ is a set of row vectors, differences between which lie in $\mathcal{K}_{\mathcal{F}}$. Similarly an element of $\mathcal{F}/K_{\mathcal{F}}$ is a set of column vectors which is parallel to $K_{\mathcal{H}}$. What does a linear functional — a real valued linear function — on $\mathcal{H}/K_{\mathcal{F}}$ look like? It is

simply a linear functional on $\mathcal{H}$ which is constant along directions which lie in $K_{\mathcal{F}}$. Every $f \in \mathcal{F}$ is a linear functional on $\mathcal{H}$ which is zero on all of $K_{\mathcal{F}}$, so every $f \in \mathcal{F}$ is a linear functional on $\mathcal{H}/K_{\mathcal{F}}$. Note that $f_1, f_2 \in \mathcal{F}$ represent the same functional on $\mathcal{H}$ — and thus on $\mathcal{H}/K_{\mathcal{F}}$ — if and only if f_1 and f_2 differ by an element of $K_{\mathcal{H}}$. So all of the column vectors in any $e \in \mathcal{F}/K_{\mathcal{H}}$ represent the same linear functional on $\mathcal{H}/K_{\mathcal{F}}$. So we may say that e is that functional. And if $g \in \mathcal{H}/K_{\mathcal{F}}$, we have $ge = hf$ for any $h \in \psi^{-1}(g)$ and for any $f \in \mathcal{F}$ taken by the quotient map to e. Likewise, each $g \in \mathcal{H}/K_{\mathcal{F}}$ is a unique linear functional on $\mathcal{F}/K_{\mathcal{H}}$. So $\mathcal{H}/K_{\mathcal{F}}$ and $\mathcal{F}/K_{\mathcal{H}}$ like $\mathcal{H}$ and $\mathcal{F}$, consist of elements of each other's dual spaces. But unlike elements of $\mathcal{H}$ and $\mathcal{F}$, elements of $\mathcal{H}/K_{\mathcal{F}}$ and $\mathcal{F}/K_{\mathcal{H}}$ contain no redundancy.

Lemma 4.2.2. $\mathcal{H}/K_{\mathcal{F}}$ and $\mathcal{F}/K_{\mathcal{H}}$ are each other's dual spaces.

Proof. We have seen that every $e \in \mathcal{F}/K_{\mathcal{H}}$ is a unique linear functional on $\mathcal{H}/K_{\mathcal{F}}$. Thus, to show that $\mathcal{F}/K_{\mathcal{H}}$ is the dual of $\mathcal{H}/K_{\mathcal{F}}$, we need only show that $\mathcal{F}/K_{\mathcal{H}}$ contains the entire dual space rather than a proper subspace. Similarly, we must show that $\mathcal{H}/K_{\mathcal{F}}$ contains the entire dual of $\mathcal{F}/K_{\mathcal{H}}$.

Let n be the dimension of $\mathcal{H}/K_{\mathcal{F}}$, and let n be the dimension of $\mathcal{F}/K_{\mathcal{H}}$. Let $g_1, \ldots, g_n$ be a linearly independent basis for $\mathcal{H}/K_{\mathcal{F}}$ and let $e_1, \ldots, e_m$ be a linearly independent basis for $\mathcal{F}/K_{\mathcal{H}}$. Finally, let A be the $n \times m$ matrix with entries $A_{ij} = g_i e_j$. Suppose a linear combination $a_1 g_1 + \ldots a_n g_n$ is taken to zero by all of $e_1, \ldots, e_m$. It must be taken to zero by every $e \in \mathcal{F}/K_{\mathcal{H}}$, and hence $(a_1 g_1 + \ldots a_n g_n)f = 0$ for every $f \in \mathcal{F}$. This is possible only if $a_1 g_1 + \ldots a_n g_n = 0$, so the rows of A are linearly independent and A has rank n. A similar argument shows that A has rank m, so $n = m$. Hence $\mathcal{H}/K_{\mathcal{F}}$ and $\mathcal{F}/K_{\mathcal{H}}$ have the same dimension, so neither can be a proper subspace of the other's dual space. Each must be the dual space of the other. ■

Bases We will now move on to more concrete and more readily manipulated forms of these vector spaces. In particular, we will be working with bases for subspaces of $\mathcal{H}$ and $\mathcal{F}$ that are isomorphic to $\mathcal{H}/K_{\mathcal{F}}$ and $\mathcal{F}/K_{\mathcal{H}}$, respectively. In addition, we will want to be able to refer to our basis vectors in a way that does not depend on any basis derived from a presentation. Thus we will choose basis vectors which are associated with words.

Definition 4.2.3. A *wordlist* W of length l is a finite list (ordered set) of words $w_1, \ldots, w_l \in \mathcal{X}^*$.

Given a wordlist W of length l and a GHMM $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$, we define the $l \times |V|$ matrix H as follows: the ith row of H is $N(\pi T^{w_i})$. We call W the *history wordlist* and H the *history matrix*. Similarly, given a wordlist S of length m, which we will call the *future wordlist*, we define the *future matrix* F to be the $|V| \times m$ matrix whose ith column is $T^{s_i}\vec{1}$. Note that H and F are functions of W and S, respectively. We will sometimes write $H(W)$ and $F(S)$ to avoid ambiguity. We will use $\overline{H}$ and $\overline{F}$ to denote the span of the rows of H and the columns of F, respectively.

Because we want $\overline{H}$ to contain a representation of every process state, we are interested in wordlists which induce a sufficiently large basis.

Definition 4.2.4. A history wordlist W is *sufficient* for a given GHMM, or simply sufficient if the span of the rows of $H(W)$ satisfies

$$\text{span}\left\{\overline{H} \cup K_{\mathcal{F}}\right\} = \mathcal{H}. \tag{4.35}$$

Similarly, a future wordlist S is sufficient if the span of the columns of $F(S)$ satisfies $\text{span}\left\{\overline{F} \cup K_{\mathcal{H}}\right\} = \mathcal{F}$.

This definition means that a history wordlist, for example, is sufficient if the rows of H, mapped into $\mathcal{H}/K_{\mathcal{F}}$ by the quotient map ψ, form a basis for $\mathcal{H}/K_{\mathcal{F}}$.

It is possible for a sufficient wordlist to contain words which are not needed for sufficiency. As we will see shortly, removing such words is desirable.

Definition 4.2.5. A history wordlist is *minimal* if it is sufficient and it has length $l = \dim\{\mathcal{H}\} - \dim\{K_{\mathcal{F}}\}$. Similarly, a future wordlist is minimal if it is sufficient and it has length $m = \dim\{\mathcal{F}\} - \dim\{K_{\mathcal{H}}\}$.

If W and S are minimal history and future wordlists, then $\overline{H}$ and $K_{\mathcal{F}}$ are complimentary subspaces of $\mathcal{H}$, and $\overline{F}$ and $K_{\mathcal{H}}$ are complimentary subspaces of $\mathcal{F}$.

Proposition 4.2.6. If W is a minimal history wordlist, then $\overline{H}$ is isomorphic to $\mathcal{H}/K_{\mathcal{F}}$. Similarly, if S is a minimal future wordlist then $\overline{F}$ is isomorphic to $\mathcal{F}/K_{\mathcal{H}}$. In both cases, the restriction of the quotient map — to $\overline{H}$ and $\overline{F}$ as appropriate — is a suitable isomorphism.

Proof. If W is a wordlist of length l, then $\dim\{\overline{H}\} \leq l$. But if W is sufficient, then $\overline{H}$ and $K_{\mathcal{F}}$ together span $\mathcal{H}$, so

$$\dim\{\overline{H}\} + \dim\{K_{\mathcal{F}}\} \geq \dim\{\mathcal{H}\}. \tag{4.36}$$

If W is also minimal, this implies $\dim\{\overline{H}\} \geq l$, so we have $\dim\{\overline{H}\} = l$ and

$$\dim\{\overline{H}\} + \dim\{K_{\mathcal{F}}\} = \dim\{\mathcal{H}\}. \tag{4.37}$$

Thus the rows of H must be linearly independent. Further, $\overline{H}$ must be independent of $K_{\mathcal{F}}$, so we can write

$$\mathcal{H} = \overline{H} \oplus K_{\mathcal{F}}. \tag{4.38}$$

Thus $\overline{H}$ is complementary to $K_{\mathcal{F}}$, and the restriction of the quotient map to $\overline{H}$ is an isomorphism between $\overline{H}$ and $\mathcal{H}/K_{\mathcal{F}}$. A similar proof holds for S, $\overline{F}$, and $\mathcal{F}/K_{\mathcal{H}}$.■

Proposition 4.2.7. If W and S are minimal wordlists, then

1. $|W| = |S|$,
2. rank(H) = rank(F),
3. dim$(\overline{H})$ = dim$(\overline{F})$, and
4. $\overline{H}$ and $\overline{F}$ are each other's dual spaces.

Proof. $\overline{H}$ is isomorphic to $\mathcal{H}/K_{\mathcal{F}}$, and $\overline{F}$ is isomorphic to $\mathcal{F}/K_{\mathcal{H}}$. Lemma 4.2.2 tells us that $\mathcal{H}/K_{\mathcal{F}}$ and $\mathcal{F}/K_{\mathcal{H}}$ are dual to each other, and we know that these isomorphisms preserve the products of elements. This $\overline{H}$ and $\overline{F}$ must be each other's dual spaces, which proves (4). This, in turn, implies (1), (2), and (3).■

We began this section by defining the vector spaces $\mathcal{H}$ and $\mathcal{F}$ so that elements of $\mathcal{H}$ represent conditional distributions on the future, and elements of $\mathcal{F}$ represent conditional distributions on the past. But $\mathcal{H}$ and $\mathcal{F}$ contain subspaces of redundancy, $K_{\mathcal{F}}$ and $K_{\mathcal{H}}$. Now, if W and S are sufficient, we have vector spaces $\overline{H}$ and $\overline{F}$, the elements of which still encode the same set of conditional distributions, and we have bases H and F for $\overline{H}$ and $\overline{F}$, respectively. In addition, if W and S are minimal, $\overline{H}$ and $\overline{F}$ contain no redundancy and H and F are linearly independent bases.

Note that we have now shown that if our history wordlist is minimal, $\overline{H}$ is isomorphic to $\mathcal{U}$. In particular, we have a natural map from one to the other: If $\mathbf{A} \in \mathcal{U}$, then there

is a unique $h \in \overline{H}$ such that for all $s \in \mathcal{X}^*$, we have $\mathbf{A}(s) = hT^s\vec{1}$. That is, if h is a mixed state, it is the mixed state representation for $\mathbf{A}$. (This h may be found by the techniques used in the proof of lemma 4.2.1, which involve building a basis of process states and a corresponding basis of mixed states induced by the same words.) Thus $\overline{H}$ may be thought of as a version of $\mathcal{U}$ consisting of tangible vectors of real numbers rather than the more abstract signed measures on an infinite sequence space.

As the following proposition establishes, if W and S are sufficient, then the matrices H and F, like the vector spaces $\mathcal{H}$ and $\mathcal{F}$, can tell us whether or not a distinction is independent of the past or of the future. And if W and S are minimal, then H and F contain no redundancy, in the sense that no row vector in the span of H is independent of the future, and likewise for F. We will work with minimal wordlists whenever we can because the absence of redundancy makes it easier to determine whether or not two processes states are distinct.

Proposition 4.2.8. If W and S are sufficient, $f \in \mathcal{F}$, and $h \in \mathcal{H}$, then

1. $f \in K_{\mathcal{H}}$ if and only if $Hf = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$, and
2. $h \in K_{\mathcal{F}}$ if and only if $hF = (0, \ldots, 0)$.

Proof. If $f \in K_{\mathcal{H}}$ and h is a row of H, then $h \in \mathcal{H}$, so $hf = 0$. If $f \notin K_{\mathcal{H}}$, then there is some $h \in \mathcal{H}$ such that $hf \neq 0$. Any vector in $\mathcal{H}$ can be written as $h = cH + k$, a linear combination of the rows of H plus some $k \in K_{\mathcal{F}}$. Thus we have

$$0 \neq hf = cHf + kf, \tag{4.39}$$

where we know that $kf = 0$. So this becomes $0 \neq cHf$, which implies $Hf \neq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$. This proves (1), and a virtually identical argument proves (2)■

Recall that each row of H is associated with a particular word: the ith row of H is the mixed state $N(\pi T^{w_i})$, where w_i is the ith element of W. Similarly, the jth column of F is $T^{s_j}\vec{1}$, where s_j is the jth element of S. Thus HF is the matrix of conditional probabilities given by

$$(HF)_{ij} = N(\pi T^{w_i})T^{s_j}\vec{1} = \mathbf{P}(s_j|w_i). \tag{4.40}$$

Also, because W and S have the same length, HF is square.

Corollary 4.2.9. If W and S are minimal, then HF is a nonsingular matrix. Conversely, if W and S are sufficient and one or both of W and S is not minimal, then HF is singular.

Proof. Suppose v is a row vector such that $vHF = (0, \ldots, 0)$. Then $(vH)F = (0, \ldots, 0)$, so $vH \in K_{\mathcal{F}}$. But $vH \in \overline{H}$, and since W is minimal, $\overline{H} \cap K_{\mathcal{F}} = \{(0, \ldots, 0)\}$. Thus we have $vH = (0, \ldots, 0)$. The rows of H are linearly independent, so we must have $v = (0, \ldots, 0)$. That is, if a linear combination of the rows of HF is the zero row, then all of the coefficients of the linear combination are zero. Hence the rows of HF are linearly independent. And HF is square, so it is nonsingular.

Conversely, assume that W is not minimal. The case in which S is not minimal may be treated similarly. Let W' be a minimal wordlist which is a subset of W, and let H' be the history matrix induced by W'. Now we consider two cases. In the first, the rows of H span the same space as the rows of H'. This means that the dimension of $\overline{H}$ must be less than the number rows in H. Thus the rows of H must be linearly dependent, which implies that the rows of HF are linearly dependent. In the second, $\overline{H}$ is larger than the span of the rows of H'. But the span of the rows of H' is isomorphic to $\mathcal{H}/K_{\mathcal{F}}$. Any larger subspace of $\mathcal{H}$ must contain nonzero vectors which lie in $K_{\mathcal{F}}$ and which are sent to zero by F. Thus the row span of H is reduced by multiplication by F, which means that the rows of HF are linearly dependent.■

Wordlists In the next section, we will use H and F extensively, as they will be our primary tools for solving the equivalence and minimization problems. In the remainder of this section, we will discuss the construction of these matrices and the construction of the wordlists on which they depend.

The next lemma is a minor fact which we will need in section 4.3.

Lemma 4.2.10. If S is sufficient and $h \in K_{\mathcal{F}}$, then for any k in the alphabet $\mathcal{X}$, $hT^k \in K_{\mathcal{F}}$.

Proof. Our assumption that $h \in K_{\mathcal{F}}$ implies that for all words s, $hT^s\vec{1} = 0$. This means that for any k, for all words s, $hT^{ks}\vec{1} = 0$, simply because ks is a word. Thus we have $\big(hT^k\big)T^s\vec{1} = 0$ for all words s, which implies that $hT^k \in K_{\mathcal{F}}$.■

Two facts make the construction of sufficient wordlists feasible. First, $\mathcal{H}$ is a vector space spanned by the mixed states, which are vectors of length $|V|$, so a basis for $\mathcal{H}$ — or any subspace of $\mathcal{H}$ — need not have more than $|V|$ elements. Similarly, a basis for $\mathcal{F}$ need not have more than $|V|$ elements. So wordlists never need to be longer than $|V|$ in order to be sufficient. Second, we have the following fact of linear algebra.

Lemma 4.2.11. If $u \in$ span$\{v_1, \ldots, v_n\}$, and A is any matrix such that the product uA is defined, then $uA \in$ span$\{v_1A, \ldots, v_nA\}$.

Proof. There must exist numbers $c_1, \ldots, c_n$ such that $u = c_1v_1 + \ldots + c_nv_n$. But then $uA = c_1v_1A + \ldots + c_nv_nA$.■

We construct sufficient wordlists using the following algorithm, which is used for a somewhat different purpose in [8].

Algorithm 4.2.12. Let Q be a queue — a first-in, first-out list — Q of words, and let and let W be a list of words. Queue Q will store a list of words which the algorithm has determined it must examine, and W will store the developing wordlist.

1. Initialize Q to contain only the word λ, and initialize W to be empty.
2. Take a word z from the tail of Q and test whether of not $N(\pi T^z)$ lies in

$$\text{span}\{N(\pi T^w) | w \in W\} = \text{span}\{\text{rows of } H(W)\}. \tag{4.41}$$

3. If it does, discard z and skip forward to step 6. (Otherwise, continue with step 4.)
4. Add z to the wordlist W.
5. For each $x \in \mathcal{X}$, add the word zx to the head of Q.
6. If Q is not empty, go back to step 2.
7. Stop. Q is empty, and W contains the completed wordlist.

Algorithm 4.2.12 builds only history wordlists, but a virtually identical algorithm builds future wordlists.

Proposition 4.2.13. The wordlists constructed by algorithm 4.2.12 are sufficient. In fact, for these wordlists, $\mathcal{H} = \overline{H}$.

Proof. We need to show that if $h \in \mathcal{H}$, then $h \in \overline{H}$. Every $h \in \mathcal{H}$ is a linear combination of vectors of the form $N(\pi T^w)$ for some $w \in \mathcal{X}^*$, so it will suffice to consider vectors

of that form. Suppose there exists a word y such that $N(\pi T^y) \notin \overline{H}$. Let z be the longest prefix of y such that $N(\pi T^z) \in \overline{H}$. There is at least one such prefix, namely λ. Let x be the symbol which follows z in y — that is, choose $x \in \mathcal{X}$ so that zx is a prefix of y. Thus $N(\pi T^{zx})$ is not in $\overline{H}$.

Now, we have chosen z so that $N(\pi T^z)$ is in $\overline{H} = \text{span}\{N(\pi T^w)|w \in W\}$, but $N(\pi T^{zx})$ is not. By lemma 4.2.11, we know that $N(\pi T^z)T^x$ is in the span of $\{N(\pi T^w)T^x|w \in W\}$, or equivalently,

$$N(\pi T^{zx}) \in \text{span}\{N(\pi T^{wx})|w \in W\}. \tag{4.42}$$

Because the algorithm added each w to W, we know that it added the words wx to the queue. Thus the vectors $N(\pi T^{wx})$ were subsequently tested against the developing basis and added to the basis if it did not already span them. Hence we know that for all $w \in W$, the vectors$N(\pi T^{zx})$ lie in $\overline{H}$, so $N(\pi T^{zx})$ is a linear combination of vectors which are known to be in $\overline{H}$, thus it is itself in $\overline{H}$. This is a contradiction, hence no words y can exist.■

Sufficient future wordlists may be constructed by an essentially identical process.

Lemma 4.2.14. A GHMM $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ has sufficient wordlists W and S, every word of which has length less than $|V|$.

Proof. The construction we have just given has the property that if w is not added to W, then no words of the form wz can be added to W. Thus if w is in W, every prefix of w is in W as well, including the length zero word λ. This means that if a word of length l is added to a wordlist W, then W has at least one element of each of the lengths $0, 1, \ldots, l$ and thus contains at least $l + 1$ elements. Note that the wordlists constructed here never have more than $|V|$ elements because rows of H are linearly independent and span a subset of $\mathbb{R}^{|V|}$. Thus a word of length $|V|$ or more can never be added to W.

A virtually identical proof holds for S.■

Once we have sufficient wordlists, we can get minimal wordlists simply by extracting appropriate subsets. Suppose we have history and future wordlists W and S, we take $H(W)F(S)$ and delete every row which is linearly dependent on those which precede

it. The words associated with the remaining rows form a new wordlist W', and the remaining rows themselves form $H(W')F(S)$.

The rows of $H(W')F(S)$ are now linearly independent, which means the rows of $H(W')$ are independent. Moreover, because we only deleted linearly dependent rows, the row span of $H(W')F(S)$ is the same as that of $H(W)F(S)$. The properties of F are such that we can be sure $H(W')$ has the appropriate span. As we will show next, the resulting wordlist W' is minimal. The analogous construction, deleting columns instead of rows, builds us a minimal future wordlist S'.

Proposition 4.2.15. Let W and S be sufficient wordlists. If W' is a subset of W such that

1. the row span of $H(W')F(S)$ is the same as that of $H(W)F(S)$, and
2. the rows of $H(W')F(S)$ are linearly independent,

then W' is minimal. Similarly, if S' is a subset of S such that

1. the column span of $H(W)F(S')$ is the same as that of $H(W)F(S)$, and
2. the columns of $H(W)F(S')$ are linearly independent,

then S' is minimal.

Proof. As usual, we have separate statements about the past and the future, and we will only prove the one about the past, as essentially the same argument will serve for the future. We will not need to refer to $F(S')$, and will use $F = F(S)$.

We will first show that W' is sufficient, that is, that the rows of $H(W')$, together with $K_{\mathcal{F}}$, span $\mathcal{H}$.

Let $h_1, \ldots, h_n$ be the rows of $H(W)$. If $h \in \mathcal{H}$, then there exists a vector $a = (a_1, \ldots, a_n)$ such that

$$\begin{aligned} h &= a_1 h_1 + \ldots + a_n h_n + k \\ &= aH(W) + k \end{aligned} \tag{4.43}$$

for some $k \in K_{\mathcal{F}}$. Multiplying on the right by F, we have $hF = aH(W)F + kF$. We know that $kF = 0$, so hF is in the span of the rows of $H(W)F$. But the rows of $H(W')F$ have the same span, so there must exist some vector c such that $hF = cH(W')F$. This is equivalent to $(h - cH(W'))F = 0$, which means that $h - cH(W')$ is in $K_{\mathcal{F}}$. Thus

$h = cH(W') + k'$ for some $k' \in K_{\mathcal{F}}$. Thus $K_{\mathcal{F}}$ and the rows of $H(W')$ span $\mathcal{H}$, so W' is sufficient.

Showing that W' is minimal is now trivial. The rows of $H(W')F$ are linearly independent and multiplying by F cannot eliminate any linear dependency which is in the rows of $H(W')$. So the rows of $H(W')$ are linearly independent, and W' is minimal.■

In this section, we have studied the relationship between the past and the future in terms of linear algebra. As part of this study, we introduced and defined H and F, and showed how to construct them and their associated wordlists. Next, we will begin to use them to identify and eliminate redundancy from GHMM presentations.

4.3 Equivalence and Minimization of GHMMs

This section addresses two problems. The first of these, which we discussed in section 4.1, is the *identifiability* problem: When are two GHMMs equivalent? That is, when do two GHMMs define the same process? The second is the *minimization* problem: Given a GHMM, what GHMM is as small as possible — that is, has as few presentation states as is possible — but is equivalent to the given one? These two questions are closely related, and have been studied for HMMs for some time. The identifiability problem is the more famous of the two and was posed by Blackwell and Koopmans in 1957 and solved by Ito et al in 1992, by a methods similar to the one presented here.[2,7] The minimization problem is nearly solved in the same paper, and was completed by Vijay Balasubramanian.[8] The details of the method presented here are the work of the author. Notably, the standard presentation, which is important here and again in chapter 5, does not appear in any previous paper, though related presentations have appeared before beginning with [20].

Suppose we have a GHMM presentation $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ for a process $\mathcal{P}$ and wordlists W and S such that H and F are invertible. (This can only happen if there is no redundancy, that is, if $K_{\mathcal{F}}$ and $K_{\mathcal{H}}$ are trivial.) We are now going to define a sort of a canonical presentation for $\mathcal{P}$, which we will call the *standard presentation* for $\mathcal{P}$, which in this case is conjugate to $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ by H.

If we conjugate T^k by H, we get $B^k = HT^kH^{-1}$, and if we write H^{-1} as $F(HF)^{-1}$, we have

$$B^k = HT^kF(HF)^{-1}. \tag{4.44}$$

We will call the matrices B^k the *standard transition matrices* for the process $\mathcal{P}$ given W and S. As we know, each $(HF)_{ij}$ is simply $\mathbf{P}(s_j|w_i)$. Similarly, for all $k \in \mathcal{X}$ and for all $i, j \in V$,

$$\begin{aligned} \left(HT^kF\right)_{ij} &= N(\pi T^{w_i})T^kT^{s_j}\vec{1} \\ &= \mathbf{P}(ks_j|w_i). \end{aligned} \tag{4.45}$$

In words, HT^kF exhibits the action of the map T^k in terms of the bases — the rows of H and the columns of F — we have developed for the past and the future. And both $(HF)^{-1}$ and HT^kF are entirely determined by probabilities of words — by the process, rather than by the presentation. This fact is key to the algorithms of this section and the next chapter.

The same is true of the initial vector

$$\gamma = \pi H^{-1} = \pi F(HF)^{-1}, \tag{4.46}$$

since $(\pi F)_j = \mathbf{P}(s_j)$. We will refer to γ as the *standard initial vector* for $\mathcal{P}$ given W and S. Thus γ and B^k depend only on the wordlists and the process, and not on the GHMMs themselves.

The following definition assembles the standard transition matrices and the standard initial vector into a presentation. We may choose any set of size $|V|$ as the set of presentation states. It will be convenient to choose W because, as we will see shortly, the presentation states are associated with the words in W. Also, recall corollary 4.2.9, which tells us that HF is invertible.

Definition 4.3.1. If $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ is a GHMM presentation for the process $\mathcal{P}$ and W and S are wordlists such that the matrix HF is invertible, then we define the *standard presentation* for $\mathcal{P}$ given W and S to be

$$\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right), \tag{4.47}$$

where $B^k = HT^kF(HF)^{-1}$ and $\gamma = \pi H^{-1} = \pi F(HF)^{-1}$.

The following result establishes that if H and F are invertible, the standard presentation for $\mathcal{P}$ is in fact a presentation, and that it is a presentation for $\mathcal{P}$. In 4.3.10, we will establish that the standard presentation is a presentation for $\mathcal{P}$whenever HF is invertible, that is, whenever the standard presentation is defined.

Lemma 4.3.2. Let $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ be a GHMM presentation for the process $\mathcal{P}$ and let W and S be wordlists such that the matrices H and F are invertible. Then the standard presentation for $\mathcal{P}$ given W and S is a GHMM presentation for $\mathcal{P}$.

Proof. The standard presentation $\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right)$ is conjugate to $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ by H. Given this fact, proposition 4.1.7 tells us that $\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right)$ is a GHMM and that it is equivalent to $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$.■

Suppose we have a second GHMM $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ which is equivalent to our first GHMM, $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$. If we take its history and future matrices H' and F' with respect to the same wordlists W and S, then we must have $H'F' = HF$, $H'U^kF' = HT^kF$, and $\tau F' = \pi F$. The for both presentations generate the same γ and the same set of B^ks and so they generate the same standard presentation. In this sense, the standard presentation plays a role similar to that of a canonical form. But because the standard presentation depends on the wordlists W and S, there is no single, absolute standard presentation. This is why we call it the *standard* presentation and not the *canonical* presentation.

In other words, suppose we have two equivalent GHMMs, and we have wordlists such that both GHMMs' history and future matrices are invertible. Then they must produce the same γ and B^k, and therefore they must both be conjugate to the standard presentation. We know that if two GHMMs are both conjugate to a third GHMM, then they must be equivalent. So using standard presentations shows promise of resolving the identifiability problem. The approach above for constructing the standard transition matrices will not work in general, because the assumption that H and F are invertible may fail. However, there is a generalization of the standard presentation which we can always construct, and so we can give a new solution to the identifiability problem using this generalization.

Theorem 4.3.3. Suppose we are given two GHMMs $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and

$\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$, and let W and S be history and future wordlists which are sufficient for both of them. Let $\mathbf{P}$ be the probability measure induced by $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and let $\mathbf{Q}$ be that induced by $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$. The GHMMs are equivalent — that is, $\mathbf{P} = \mathbf{Q}$ — if and only if all of the following hold:

1. for all $w_i \in W$ and $s_j \in S$, $\mathbf{P}(s_j|w_i) = \mathbf{Q}(s_j|w_i)$,
2. for all $w_i \in W$, $s_j \in S$, and $k \in \mathcal{X}$, $\mathbf{P}(ks_j|w_i) = \mathbf{Q}(ks_j|w_i)$, and
3. for all $s_j \in S$, $\mathbf{P}(s_j) = \mathbf{Q}(s_j)$.

We will prove this by showing that the probability of any word is determined by these few probabilities, these few conditional probabilities, and the information that W and S are sufficient. Thus the entire process is determined by these same few pieces of information. At this point, it is worth recalling Bayes rule — $\mathbf{P}(s|w) = \mathbf{P}(ws)/\mathbf{P}(w)$ — because it tells us that we can compute the necessary conditional probabilities from the (non-conditional) probabilities of all words w_i, s_j, w_is_j, and w_iks_j for all $w_i \in W$, $s_j \in S$, and $k \in \mathcal{X}$. We can ignore the possibility that $\mathbf{P}(w_i) = 0$ for some i, and hence that $\mathbf{P}(s_j|w_i)$ will not be well-defined, because the row of H corresponding to that w_i must lie in $K_{\mathcal{F}}$. Such a row is never needed in the basis.

The proof itself, which begins on page 79, uses a number of lemmas, most of which are proved by calculation and use of the properties of H and F. These lemmas develop a generalization of the standard presentation. Note that we are using W as a set of presentation states. Recall that we concluded in chapter 3 that the role of presentation states was to serve as basis vectors for the space containing the mixed states. Here, we use this the other way, and having chosen a basis for the mixed states, we will use the elements of that basis as presentation states. Arguably, the presentation states should be labeled with the mixed states, rather than the strings in W which induce those mixed states, but we will use the strings themselves for brevity and clarity.

Whenever two GHMMs share an alphabet†, it is always possible to find wordlists which are sufficient for both of them. If W_1 and W_2 are sufficient for $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$, respectively, then the wordlist $W = W_1 \cup W_2$, with words in any fixed order, will be sufficient for both $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$.

†Two GHMMs which have distinct alphabets always represent different processes. In some applications, however, it may be desirable to map one alphabet onto another so as to sidestep this fact.

When W and S are not minimal, and thus HF is not invertible, we can no longer define the standard transition matrices B^k. It is still possible, however, to define a form of the standard presentation, though we can no longer define it as simply as we can if W and S are minimal.

Definition 4.3.4. Given a GHMM presentation $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ of a process $\mathcal{P}$ and sufficient wordlists W and S, we say that $\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right)$ is a *quasi-presentation* if

1. for each $k \in \mathcal{X}$, B^k satisfies $B^k HF = HT^k F$, and
2. γ satisfies $\gamma HF = \pi F$.

We refer to B^k and γ as a *quasi-transition matrix* and a *quasi-initial vector* respectively.

If HF is invertible, then there is a unique quasi-presentation for a process $\mathcal{P}$ given W and S, and it is the standard presentation. Otherwise, quasi-presentations are not unique. Lemma 4.3.5 shows that quasi-presentations exist.

Lemma 4.3.5. If W and S are sufficient, then for each k there exists a quasi-transition matrix B^k such that

$$B^k HF = HT^k F, \tag{4.48}$$

and there exists a quasi-initial vector γ such that $\gamma HF = \pi F$.

Proof. Let $h_i = N(\pi T^{w_i})$ be the ith row of H. Then the ith row of HT^k is $h_i T^k = N(\pi T^{w_i}) T^k$, which must lie in $\mathcal{H}$. The rows of H and elements of $K_{\mathcal{F}}$ span $\mathcal{H}$, so there is some $v \in K_{\mathcal{F}}$ and some $a \in \mathbb{R}^n$ such that $h_i T^k = aH + v$. This means that $h_i T^k F = aHF$. Let the ith row of B^k be a.

Likewise, π is in $\mathcal{H}$, so we can find $\gamma \in \mathbb{R}^n$ and $v \in K_{\mathcal{F}}$ such that $\gamma H + v = \pi$, and hence $\gamma HF = \pi F$.■

The next lemma depends on the normalization of the rows of H. In fact it is the reason we defined H — and indeed, $\mathcal{H}$ — using $N(\pi T^w)$ instead of the simpler πT^w.

Lemma 4.3.6. The matrix $\sum_{k \in \mathcal{X}} B^k$ and the vector γ are both unit-sum.

Proof. Because $\vec{1}$ is an element of $\mathcal{F}$, there exists an $a \in \mathbb{R}^n$ and an $u \in K_{\mathcal{H}}$ such that $\vec{1} = Fa + u$. Thus $H\vec{1} = HFa$, and since the rows of H are mixed states and so must

be unit-sum, we have $\vec{1} = HFa$. Similarly, for all $k \in \mathcal{X}$, we can write

$$HT^k\vec{1} = HT^kFa + HT^ku \tag{4.49}$$

But each row of HT^k is in $\mathcal{H}$, so $HT^ku = 0$ and we have $HT^k\vec{1} = HT^kFa$. Now, $\sum_k T^k$ is unit-sum, so if we sum on k, this becomes $\vec{1} = H\sum_k T^kFa$.

Summing 4.48 on k gives us

$$\left(\sum_{k\in\mathcal{X}} B^k\right)HF = H\left(\sum_{k\in\mathcal{X}} T^k\right)F. \tag{4.50}$$

If we multiply on the right by a, we can substitute $\vec{1}$ for HFa and for $H\sum_k T^kFa$, and we have $\left(\sum_k B^k\right)\vec{1} = \vec{1}$.

Similarly, $\gamma HF = \pi F$ becomes $\gamma HFa = \pi Fa$, which in turn becomes $\gamma\vec{1} = 1$.■

Lemma 4.3.6 establishes that $\left(W, \mathcal{X}, \{B^k\}, \gamma\right)$ satisfies conditions 1 and 2 of the three conditions of definition 4.1.2, which defines a Proto-GHMM. The next lemma tells us that it may fail to satisfy the final condition — $\gamma\sum_k B^k$ may differ from γ, but only by a vector in $K_{\mathcal{F}}$.

Lemma 4.3.7. For some vector r such that $rHF = 0$, γ satisfies $\gamma\sum_k B^k = \gamma + r$.

Proof. Because γ is known to satisfy $\gamma HF = \pi F$, there is a $v \in K_{\mathcal{F}}$ such that $\gamma H = \pi + v$. Thus,

$$\begin{aligned}\gamma H\sum_k T^k &= \pi\sum_k T^k + v\sum_k T^k \\ &= \pi + v'\end{aligned} \tag{4.51}$$

for some $v' \in K_{\mathcal{F}}$. Similarly, 4.50 tells us that

$$H\sum_k T^k = \sum_k B^kH + A \tag{4.52}$$

for some matrix A, all rows of which lie in $K_{\mathcal{F}}$. When we substitute the right-hand side of equation 4.52 into equation 4.51, we have $\gamma\sum_k B^kH = \pi + v' - \gamma A$, from which the substitution of $\gamma H - v$ for π gives us

$$\gamma\sum_k B^kH = \gamma H + v' - \gamma A - v. \tag{4.53}$$

Multiplying by F, we have

$$\gamma \sum_k B^k HF = \gamma HF. \tag{4.54}$$

Thus, there is some row vector r such that $rHF = 0$ and $\gamma \sum_k B^k = \gamma + r$.■

We would like lemma 4.3.7 to have established that $\gamma \sum_k B^k = \gamma$, because that would have completed the verification that $\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right)$ is a Proto-GHMM. However, this is not always the case — a quasi-presentation is not always a Proto-GHMM. If HF is invertible — that is, if W and S are minimal — then $\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right)$ is the standard presentation. In this case, we do have a Proto-GHMM, which we may prove by multiplying equation 4.54 by $(HF)^{-1}$. We will soon show that the standard presentation is, in fact, a GHMM equivalent to $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$, and we will use this fact later in this section when we address the minimization problem.

In the next lemma, which is the key step in the proof of theorem 4.3.3, we establish that the quasi-initial vector γ and the quasi-presentation matrices B^k can reproduce the probabilities of words given by the initial vector π and the transition matrices T^k. Here, we begin using the quasi-transition matrices as transition matrices, with the convention that if $x = x_1 \ldots x_n$ is a word, then $B^x = B^{x_1} \cdots B^{x_n}$.

Lemma 4.3.8. For all $x \in \mathcal{X}^*$,

1. $HT^x F = B^x HF$,
2. $\pi T^x F = \gamma B^x HF$, and
3. $\pi T^x \vec{1} = \gamma B^x \vec{1}$.

Proof. We will prove by induction on the length of x. If the length is one, then 1 is equivalent to $B^k HF = HT^k F$, which B^k satisfies by definition.

If x has length greater than one, then we can write $x = yk$ for $k \in \mathcal{X}$ and y the prefix of x with length $|x| - 1$. We will assume, as our induction assumption, that $HT^y F = B^y HF$. This means that there is some matrix A such that

$$HT^y = B^y H + A, \tag{4.55}$$

and all rows of A lie in $K_{\mathcal{F}}$. Now, multiply on the right by T^k and we have

$$HT^y T^k = B^y HT^k + AT^k. \tag{4.56}$$

Because B^k is defined to satisfy $B^k HF = HT^k F$, we know that $HT^k = B^k H + C$ for some matrix C, all rows of which lie in $K_{\mathcal{F}}$. Making this substitution for HT^k on the right-hand side of 4.56 and writing $C' = C + AT^k$, we have

$$HT^y T^k = B^y B^k H + C'. \tag{4.57}$$

Note that $C'F$ is a matrix of zeros. Thus, if we multiply equation 4.57 by F, we have

$$HT^x F = B^x HF, \tag{4.58}$$

which proves 1.

Note that we have shown that

$$HT^x = B^x H + A \tag{4.59}$$

for some matrix A such that AF is a matrix of zeros. Multiplying by γ and then substituting $\pi + v$ for γH gives us

$$\pi T^x + vT^x = \gamma B^x H + \gamma A, \tag{4.60}$$

for some $v \in K_{\mathcal{F}}$. Note that both vT^x and γA lie in $K_{\mathcal{F}}$. Now, if we multiply by F, we get

$$\pi T^x F = \gamma B^x HF, \tag{4.61}$$

thus proving 2.

And finally, if we multiply equation 4.60 by $\vec{1}$, we have

$$\pi T^x \vec{1} = \gamma B^x H\vec{1}. \tag{4.62}$$

Because H is a unit-sum matrix, equation 4.62 proves 3.■

We have now defined quasi-presentations, proven that they exist, and proven that they determine the probabilities of all words. Having done so, we are ready to prove theorem 4.3.3.

Proof of Theorem 4.3.3. If two GHMMs $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ are equivalent, then they agree on the probabilities of all words, and thus an all conditional

probabilities. The interesting part of the theorem is that if W and S are sufficient for both $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$, and if these two GHMMs agree on the probabilities $\mathbf{P}(s|w)$, $\mathbf{P}(ks|w)$, and $\mathbf{P}(s)$ for all $w \in W$, $s \in S$, and $k \in \mathcal{X}$, then $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ are equivalent. We will prove this by constructing a quasi-presentation for $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and showing that it is also a quasi-presentation for $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$.

Let H_1 and F_1 be the history and future matrices for $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and let H_2 and F_2 be the history and future matrices for $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$. Let us recall the hypotheses of this theorem.

1. For all $w_i \in W$ and $s_j \in S$, $\mathbf{P}(s_j|w_i) = \mathbf{Q}(s_j|w_i)$,
2. For all $w_i \in W$, $s_j \in S$, and $k \in \mathcal{X}$, $\mathbf{P}(ks_j|w_i) = \mathbf{Q}(ks_j|w_i)$, and
3. For all $s_j \in S$, $\mathbf{P}(s_j) = \mathbf{Q}(s_j)$.

First, $\mathbf{P}(s_j|w_i) = (H_1F_1)_{ij}$, and $\mathbf{Q}(s_j|w_i) = (H_2F_2)_{ij}$, so 1 is equivalent to $H_1F_1 = H_2F_2$. Second, $\mathbf{P}(ks_j|w_i) = \left(H_1T^kF_1\right)_{ij}$, and $\mathbf{Q}(ks_j|w_i) = \left(H_2U^kF_2\right)_{ij}$, so 2 can be written as follows: for all k, $H_1T^kF_1 = H_2U^kF_2$. Last, $\mathbf{P}(s_j) = (\pi F_1)_j$, and $\mathbf{Q}(s_j) = (\tau F_2)_j$, so 3 becomes $\pi F_1 = \tau F_2$. Thus, what we need to show is that these three facts — $H_1F_1 = H_2F_2$, $H_1T^kF_1 = H_2U^kF_2$ for all k, and $\pi F_1 = \tau F_2$ — together imply that $\mathbf{P}(x) = \mathbf{Q}(x)$ for any $x \in \mathcal{X}^*$, where $\mathbf{P}(x) = \pi T^x\vec{1}$ and $\mathbf{Q}(x) = \tau U^k\vec{1}$. Let γ be any solution to $\gamma H_1F_1 = \pi F_1$, and for all k, let B^k be any solution to $B^kH_1F_1 = H_1T^kF_1$. Then $\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right)$ is a quasi-presentation for the process represented by $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$. Lemma 4.3.8 now tells us that for any $x \in \mathcal{X}^*$ $\pi T^x\vec{1} = \gamma B^x\vec{1}$.

But we know that $H_1F_1 = H_2F_2$ and $\pi F_1 = \tau F_2$, so γ satisfies $\gamma H_2F_2 = \tau F_2$. Similarly, each B^k satisfies $B^kH_2F_2 = H_2U^kF_2$. Thus, $\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right)$ is also a quasi-presentation for $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$. So lemma 4.3.8 tells us that for all x, $\tau U^x\vec{1} = \gamma B^x\vec{1}$, and therefore that for all x, $\tau U^x\vec{1} = \pi T^x\vec{1}$.■

Corollary 4.3.9. If $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ are two GHMMs which assign the same probabilities to all words of length less than $2n$, neither of which has more than n states, then they are equivalent.

Proof. Lemma 4.2.14 tells us that we can find wordlists W and S in which all of the words have length at most $n-1$. For such wordlists, all words of the form w_i, s_j, $w_i s_j$, and $w_i k s_j$ have lengths less than $2n$, so $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ must agree on all $\mathbf{P}(s_j|w_i) = \mathbf{P}(w_i s_j)/\mathbf{P}(w_i)$, $\mathbf{P}(k s_j|w_i) = \mathbf{P}(w_i k s_j)/\mathbf{P}(w_i)$, and $\mathbf{P}(s_j)$. Thus, by theorem 4.3.3, they are equivalent.■

With this machinery in hand, we can resolve the minimization problem.

Theorem 4.3.10. Given a GHMM $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$, let $\mathbf{P}$ be the process it represents, and let W and S be minimal wordlists. Then the standard presentation $\left(W, \mathcal{X}, \left\{B^k\right\}, \gamma\right)$ for given W and S is a GHMM, and it is equivalent to $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$. Furthermore, no GHMM exists with fewer than $|W|$ states which is equivalent to $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$.

Proof. As noted on page 78, when HF is invertible, $\gamma = \sum_k B^k \gamma$, and the standard presentation is a Proto-GHMM. We established in lemma 4.3.8 that $\gamma B^x \vec{1} = \pi T^x \vec{1}$ for all $x \in \mathcal{X}^*$. This proves both that it is a GHMM and that it is equivalent to $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$.

Because the rows of HF are linearly independent, and these rows consist of conditional future probabilities, we know that the process states $\mathbf{P}(\cdot|w_i)$ are linearly independent. Thus the span of the reachable process states has dimension at least $|W|$. In fact, if we combine lemma 4.2.1 and proposition 4.2.6, we have proven that its dimension is exactly $|W|$. If a GHMM has fewer than $|W|$ states, then it induces a process for which the span of the reachable process states has dimension less than $|W|$, and thus it cannot be equivalent to $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$.■

We conclude this section with a result — the existence of conjugacies — which we promised in section 4.1. This was first shown — for functions of finite Markov Chains — by Gilbert [20].

Proposition 4.3.11. Let $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ be two minimal, equivalent GHMMs — that is, $|V| = |V'| = \dim(\mathcal{U})$, where $\mathcal{U}$ is the span of the reachable process states. Then there exists a matrix M such that $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ are conjugate by M.

Proof. Let W and S be minimal wordlists for $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$, and let H_1 and F_1 be the associated history and future matrices. Because W and S are minimal, $H_1 F_1$

must be invertible. We know that $|V| = \dim(\mathcal{U})$, and for minimal W we know that $|W| = \dim(\mathcal{U})$. But H_1 has $|W|$ rows and $|V|$ columns, so it must be square. Likewise, F_1 must be square and both H_1 and F_1 must be invertible.

Let H_2 and F_2 be the history and future matrices for $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$. Then the process states $\mathbf{P}(\cdot|w_i)$ form a linearly independent basis for $\mathcal{U}$, hence the rows of H_2 must be linearly independent. Thus H_1, F_1, and H_2 are full rank square matrices, and must be invertible. We also know that $H_1F_1 = H_2F_2$ is invertible, so $F_2 = H_2^{-1}H_1F_1$ must be invertible.

Now, we calculate. We have $H_2U^kF_2 = H_1T^kF_1$ and $\pi F_1 = \tau F_2$, so

$$U^k = H_2^{-1}H_1T^kF_1F_2^{-1} \tag{4.63}$$

and $\tau = \pi F_1F_2^{-1}$. And if we let $M = H_2^{-1}H_1$, M is a unit-sum matrix. And $H_1F_1 = H_2F_2$ implies $H_2^{-1}H_1F_1F_2^{-1} = I$, so $M^{-1} = F_1F_2^{-1}$. So equation 4.63 may be written $U^k = MT^kM^{-1}$. Moreover, we have $\tau = \pi F_1F_2^{-1} = \pi M^{-1}$. Thus $\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$ and $\left(V', \mathcal{X}, \left\{U^k\right\}, \tau\right)$ are conjugate.■

We began this section by defining a GHMM to be a representation of a process similar to an HMM, but with negative entries allowed in its transition matrices. We studied the vector spaces of conditional distributions on the future induced by history words and of conditional distributions on the past induced by future words. We found bases for these vector spaces in terms of these same words. These bases were instrumental in resolving the identifiability and minimization problems, and they led us to the standard presentation. We have shown that the standard presentation is a GHMM presentation for a process, and we have observed that it is determined entirely by probabilities of words. In the next chapter, we will use this fact to construct presentations directly from probabilities of words — that is, directly from the process.