Chapter 4

Faster HSMM Inference with Efficient Representations

4.1 Introduction

In this chapter we address some fundamental scaling challenges for Hidden semi-Markov Model (HSMM) inference. One reason HSMMs are not utilized nearly as often as the ubiquitous Hidden Markov Model (HMM) is that the HSMM's basic inference routines are often too computationally expensive in practical settings. The time complexity of HSMM inference scales quadratically with data sequence length while HMM inference scales only linearly. The slowdown relative to HMM inference is due to the weaker Markovianity properties in the HSMM state sequence, which require more expensive message passing algorithms. For this reason, with growing dataset sizes and richer hierarchical models requiring more data to be fit, HSMMs are often rejected even when explicit durations provide a better model.

We address this challenge in two ways. First, we study HSMMs with a particular natural family of duration distributions, the two-parameter negative binomial family, and develop a message passing algorithm for these models with complexity that scales only linearly with the sequence length. We derive the message passing algorithm in terms of a general notion of HMM embeddings, which we define in Section 4.3 and which improves upon previous work on expanded-state HMMs, as we discuss in Section 4.2. We also develop a Gibbs sampling algorithm that uses this linear-time message passing scheme. These algorithms immediately generalize to duration models that are mixtures of negative binomials, a natural duration analog of the Gaussian mixture model. Second, we give a linear time-invariant (LTI) system realization perspective that both generalizes the class of duration models for which HSMM message passing can be made efficient and illuminates the limits of such an approach.

In subsequent chapters we build on these inference methods for HSMMs with negative binomial durations to develop scalable inference algorithms for hierarchical Bayesian and Bayesian nonparametric models, including the stochastic variational inference (SVI)

methods for the HSMM and HDP-HSMM described in Chapter 5 and both Gibbs sampling and SVI algorithms for a new model described in Chapter 6.

The remainder of this chapter is organized as follows. In Section 4.2 we discuss previous work and how it relates to the framework and methods we develop. In Section 4.3, we provide our definition of HMM embeddings of HSMMs and give a general result on computing HSMM messages in terms of the embedding. In Section 4.4 we give a particular embedding for HSMMs with negative binomial durations and show how to use the embedding to construct an HSMM Gibbs sampling algorithm for which the time complexity of each iteration scales only linearly in the observation sequence length. Finally, in Section 4.5 we give a more general perspective on efficient representations for HSMM message passing in terms of LTI system realization.

■ 4.2 Related work

The work that is most closely related to the ideas we develop in this chapter is the work on expanded state HMMs (ESHMMs) [98, 97, 61, 48]. In the ESHMM framework, non-geometric state durations are captured by constructing an HMM in which several states share each observation distribution; the observations are generated from the same observation distribution while the hidden Markov chain occupies any state in the corresponding group of states, and so the effective duration distribution is modeled by the Markov dwell time for the group of states. This technique is similar to the notion of HMM embedding that we develop, but there are some key advantages to our approach, both modeling and algorithmic. First, while an ESHMM is limited to a single HMM of fixed size and transition topology, our definition of HMM embeddings as a way to compute HSMM messages allows us to perform Bayesian inference over the HSMM duration parameters and thus effectively resize the corresponding HMM embedding, as we show in Section 4.4.2. Second, another consequence of using HMM embeddings to compute HSMM messages is that the message passing recursions we develop require significantly less memory than those for the ESHMM, as we describe in Section 4.3. Finally, as we show in Section 4.5.2, our definition can be generalized to give efficient HSMM message passing algorithms for duration models that cannot be captured by ESHMMs.

Note that the HMM embedding we develop in Section 4.4 is most similar to the ESHMM Type A model [97, 61], which can model a modified negative binomial duration distribution with fixed r parameter and PMF given by

$$p(k|r,p) = \binom{k-1}{k-r}(1-p)^r p^{k-r} \qquad k = r, r+1, r+2, \ldots. \tag{4.2.1}$$

This distribution is not the standard negative binomial distribution which we use in Section 4.4. In particular, the PMF (4.2.1) has support starting at r, which can be an artificial limitation when modeling duration distributions and particularly when learning the r parameter, as we do in Section 4.4.

■ 4.3 HMM embeddings and HSMM messages

In this section, we give our definition for HMM embeddings of HSMMs and show how the HSMM messages can be computed in terms of HMM messages in the embeddings. In Section 4.4 we use these results to derive efficient inference algorithms for HSMMs with negative binomial durations, and in Section 4.5 we generalize these definitions.

As described in Chapter 3, there are standard HSMM forward and backward messages [78] analogous to those for the HMM. The forward messages (F, F^*) and backward messages (B, B^*) are defined by

$$\begin{aligned} F_{t,i} &\triangleq p(y_{1:t}, x_t = i, x_t \neq x_{t+1}) \\ &= \sum_{d=1}^{T-t-1} F^*_{t-d,i} p(d|x_{t-d} = i) p(y_{t-d:t}|x_{t-d:t} = i) \end{aligned} \tag{4.3.1}$$

$$\begin{aligned} F^*_{t,i} &\triangleq p(y_{1:t}, x_{t+1} = i|x_t \neq x_{t+1}) \\ &= \sum_{j=1}^{N} F_{t,i} p(x_{t+1} = i|x_t = j, x_t \neq x_{t+1}) \end{aligned} \tag{4.3.2}$$

$$F_{1,i} \triangleq p(x_1 = i) \tag{4.3.3}$$

$$\begin{aligned} B_{t,i} &\triangleq p(y_{t+1:T}|x_t = i, x_t \neq x_{t+1}) \\ &= \sum_{j=1}^{N} B^*_{t,j} p(x_{t+1} = j|x_t = i, x_t \neq x_{t+1}), \end{aligned} \tag{4.3.4}$$

$$\begin{aligned} B^*_{t,i} &\triangleq p(y_{t+1:T}|x_{t+1} = i, x_t \neq x_{t+1}) \\ &= \sum_{d=1}^{T-t} B_{t+d,i} p(d|x_{t+1} = i) p(y_{t+1:t+d}|x_{t+1:t+d} = i) \\ &\quad + \sum_{d=T-t+1}^{\infty} p(d|x_{t+1} = i) p(y_{t+1:T}|x_{t+1:T} = i), \end{aligned} \tag{4.3.5}$$

$$B_{T,i} \triangleq 1, \tag{4.3.6}$$

where T denotes the length of the observation sequence and N the number of states. These HSMM messages are expensive to compute because the expression for $B^*_{t,i}$, like that for $F_{t,i}$, involves summing over all possible durations, resulting in an overall time complexity of $\mathcal{O}(TN^2 + T^2N)$ compared to the HMM message passing complexity of $\mathcal{O}(TN^2)$. The HSMM algorithm's quadratic dependence on the sequence length T severely limits the settings in which HSMM inference can be performed and has led many practitioners to prefer HMMs even when geometric state durations are unnatural.

One way to interpret the HSMM messages messages is to *embed* the HSMM into a much larger HMM that encodes the same generative process [78, 54]. The HMM embedding includes the duration information in its Markov state. Here we give a more general definition of HMM embeddings than considered in previous work, and use this general definition to explore opportunities for more efficient representations.

Recall from Chapter 3 that we parameterize an HSMM on N states with a $N \times N$ transition matrix A where $A_{ij} = p(x_{t+1} = j | x_t = i, x_t \neq x_{t+1})$, initial state distribution $\pi^{(0)}$, observation parameters $\theta = \{\theta^{(i)}\}_{i=1}^N$, and duration parameters $\vartheta = \{\vartheta^{(i)}\}_{i=1}^N$.

Definition 4.3.1 (HMM embedding of an HSMM)**.** *Given an HSMM on N states with parameters $(A, \theta, \pi^{(0)}, \vartheta)$ and an observation sequence length T, an* HMM embedding *of the HSMM is an HMM on $\bar{N} = \sum_{i=1}^N \bar{N}^{(i)}$ states for some $\bar{N}^{(i)}$ with parameters $(\bar{A}, \bar{\theta}, \bar{\pi}^{(0)})$ that satisfy the following requirements:*

(1) The HMM transition matrix is of the form

$$\bar{A} = \begin{pmatrix} \bar{A}^{(1)} & & & \\ & \bar{A}^{(2)} & & \\ & & \ddots & \\ & & & \bar{A}^{(3)} \end{pmatrix} + \begin{pmatrix} | \\ \bar{b}^{(1)} \\ | \\ & \ddots & \\ & & | \\ & & \bar{b}^{(N)} \\ & & | \end{pmatrix} \begin{pmatrix} & A & \end{pmatrix} \begin{pmatrix} — \, \bar{c}^{(1)\mathsf{T}} \, — \\ & \ddots & \\ & & — \, \bar{c}^{(N)\mathsf{T}} \, — \end{pmatrix} \tag{4.3.7}$$

for some nonnegative matrices $\bar{A}^{(i)}$ of size $\bar{N}^{(i)} \times \bar{N}^{(i)}$ and nonnegative vectors $\bar{b}^{(i)}$ and $\bar{c}^{(i)}$ of length $\bar{N}^{(i)}$ for $i = 1, 2, \ldots, N$. Entries shown as blank are zero.

(2) Indexing each HMM state as (i, j) for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, \bar{N}^{(i)}$, the HMM observation parameters are $\bar{\theta}^{(i,j)} = \theta^{(i)}$ and the initial state distribution is $\bar{\pi}^{(0)}_{(i,j)} = \pi^{(0)}_i c^{(i)}_j$. We can thus associate each HSMM state i with a collection of $\bar{N}^{(i)}$ HMM states of the form (i, j), and we refer to each i as the state index *and each j as the* pseudostate index.

(3) The probability the HSMM assigns to any label sequence $x_{1:T} = (x_t)_{t=1}^T$ is equal to the total probability the HMM embedding assigns to all HMM state sequences

of the form $\bar{x}_{1:T} = ((x_t, e_t))_{t=1}^T$ *for some sequence of pseudostate indices* (e_t)*; that is, writing* $p(x_{1:T}|A, \pi^{(0)}, \theta, \vartheta)$ *for the probability the HSMM assigns to a fixed label sequence* $x_{1:T}$ *and* $p(\bar{x}_{1:T}|\bar{A}, \bar{\pi}^{(0)}, \bar{\theta})$ *for the probability that the HMM assigns to a state sequence* $\bar{x}_{1:T}$ *with* $\bar{x}_t = (x_t, e_t)$*, we require*

$$p(x_{1:T}|A, \pi^{(0)}, \theta, \vartheta) = \sum_{e_t} p(\bar{x}_{1:T}|\bar{A}, \bar{\pi}, \bar{\theta}) = \sum_{e_t} p((x_t, e_t)_{t=1}^T|\bar{A}, \bar{\pi}, \bar{\theta}) \tag{4.3.8}$$

where the sum is over all possible pseudostate index sequences.

Note that by the construction in (4.3.7), for any HMM embedding of an HSMM, each matrix

$$\begin{pmatrix} \bar{A}^{(i)} & \bar{b}^{(i)} \\ \bar{c}^{(i)\mathsf{T}} & 0 \end{pmatrix} \tag{4.3.9}$$

for $i = 1, 2, \ldots, N$ is row-stochastic. Further, again decomposing each $\bar{x}_t = (x_t, e_t)$ into an HSMM state index and a pseudostate index, note that by Definition 4.3.1 we have

$$\bar{c}_j^{(i)} = p(\bar{x}_{t+1} = (i, j)|x_{t+1} = i, x_t \neq x_{t+1}) \tag{4.3.10}$$

$$\bar{b}_j^{(i)} = p(x_t \neq x_{t+1}|\bar{x}_t = (i, j)). \tag{4.3.11}$$

Thus we refer to each $\bar{c}^{(i)}$ as the vector of *entrance probabilities* and to each $\bar{b}^{(i)}$ as the vector of *exit probabilities* for the HMM embedding pseudostates corresponding to HSMM state i. An HMM embedding captures the desired HSMM transition dynamics because $\sum_{k=1}^{\bar{N}^{(i)}} \sum_{\ell=1}^{\bar{N}^{(j)}} p(\bar{x}_{t+1} = (j, \ell)|\bar{x}_t = (i, k), x_t \neq x_{t+1}) = A_{ij}$. Finally, note that an HMM embedding models the HSMM durations in terms of the dwell times within the group of HMM states corresponding to each HSMM state, as we make clear in Proposition 4.3.3.

First, we give an example of a generic HSMM embedding from a construction given in Murphy [78] (though not in terms of our definition of HMM embeddings).

Example 4.3.2. *We can construct an HMM embedding for an arbitrary HSMM by using* $\bar{N} = TN$ *total states with* $\bar{N}^{(i)} = TN$ *and*

$$\bar{A}^{(i)} = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{pmatrix} \qquad \bar{b}^{(i)} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{4.3.12}$$

$$\bar{c}^{(i)\mathsf{T}} = \begin{pmatrix} p(d \geq T|\vartheta^{(i)}) & p(d = T-1|\vartheta^{(i)}) & \cdots & p(d=2|\vartheta^{(i)}) & p(d=1|\vartheta^{(i)}) \end{pmatrix}. \tag{4.3.13}$$

In this HMM embedding, when the group of states corresponding to HSMM state i is entered, a duration (censored to a maximum of T) is sampled from the HSMM state's duration distribution encoded in $\bar{c}^{(i)}$; if a duration d is sampled, the state with pseudostate index $T - d$ is entered. Then the pseudostate index serves as a counter, deterministically incrementing until reaching the maximum pseudostate index of T, at which point a transition occurs and a new HSMM state is sampled.

While the construction makes it impossible to sample a pseudostate corresponding to a duration greater than T, with the label sequence length fixed at T it is impossible for any duration longer than T to be represented in the HSMM label sequence. Thus this HMM embedding directly models the HSMM generative process for label sequences of length T. Note that, in a sense we make precise in Section 4.5, an embedding that depends on the sequence length T corresponds to an HSMM that does not have a finite-size realization.

Example 4.3.2 also makes clear the computational complexity of HSMM message passing and the reason it scales quadratically with T. The time complexity for message passing on a generic HMM with TN states is $\mathcal{O}(T(TN)^2)$ because at each time one must compute a matrix-vector product with a matrix of size $(TN) \times (TN)$. However, due to the structure of $\bar{A}$ each multiplication for the HMM embedding can be done in $\mathcal{O}(N^2 + TN)$ time. Indeed, this generic HSMM embedding provides a way to derive the HSMM messages in terms of the HMM messages on the embedding.

Proposition 4.3.3. *Let $(A, \theta, \pi^{(0)}, \vartheta)$ be an HSMM and let $(\bar{A}, \bar{\theta}, \bar{\pi}^{(0)})$ be a candidate HMM embedding satisfying conditions (1) and (2) of Definition 4.3.1. If for each $i = 1, 2, \ldots, N$ we have*

$$\bar{c}^{(i)\mathsf{T}} \left(\bar{A}^{(i)}\right)^{d-1} \bar{b}^{(i)} = p(d|\vartheta^{(i)}) \qquad d = 1, 2, \ldots, T-1 \tag{4.3.14}$$

$$\bar{c}^{(i)\mathsf{T}} \left(\bar{A}^{(i)}\right)^{T-1} \bar{b}^{(i)} = p(d \geq T|\vartheta^{(i)}) \tag{4.3.15}$$

then $(\bar{A}, \bar{\theta}, \bar{\pi}^{(0)})$ *is an HMM embedding of* $(A, \theta, \pi^{(0)}, \vartheta)$.

Proof. Note that the matrix expression in (4.3.14) gives the probability of absorption in d timesteps for a Markov chain with transition matrix and initial state distribution given by

$$\begin{pmatrix} \bar{A}^{(i)} & \bar{b}^{(i)} \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \bar{c}^{(i)\mathsf{T}} & 0 \end{pmatrix}, \tag{4.3.16}$$

respectively. Therefore, by Definition 4.3.1, the expression gives the probability of remaining in the pseudostates corresponding to HSMM state i for exactly d timesteps. Finally, note that, as in Example 4.3.2, the HMM embedding only needs to represent the duration distribution for durations up to the observation sequence length T. □

As we show in the next example, based on a construction from Hudson [54], HMM embeddings of an HSMM are not unique.

Example 4.3.4. *For an HSMM* $(A, \theta, \pi^{(0)}, \vartheta)$*, using Definition 4.3.1 choose*

$$\bar{A}^{(i)} = \begin{pmatrix} 0 & 1-p(d=1|\vartheta^{(i)}) & & & \\ & 0 & 1-p(d=2|d\geq 2,\vartheta^{(i)}) & & \\ & & \ddots & \ddots & \\ & & & 0 & 1-p(d=T-1|d\geq T-1,\vartheta^{(i)}) \\ & & & & 0 \end{pmatrix} \tag{4.3.17}$$

$$\bar{b}^{(i)} = \begin{pmatrix} p(d=1|\vartheta^{(i)}) \\ p(d=2|d\geq 2,\vartheta^{(i)}) \\ \vdots \\ p(d=T-1|d\geq T-1,\vartheta^{(i)}) \\ 1 \end{pmatrix} \qquad \bar{c}^{(i)\mathsf{T}} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \end{pmatrix}. \tag{4.3.18}$$

This HMM embedding also uses the pseudostate index as a duration counter, but instead of counting down until the time an HSMM transition occurs, the pseudostate here counts up the time since the previous HSMM transition.

Given any valid HMM embedding of an HSMM, we can compute the HSMM messages in terms of the HMM messages for the embedding, as we show in the following proposition.

Proposition 4.3.5 (HSMM Messages from HMM Embedding)**.** *Given an HMM embedding for some HSMM, let* $\bar{F}$ *and* $\bar{B}$ *denote the HMM messages for the embedding as in Eqs.* (2.4.6) *and* (2.4.7) *of Section 2.4, so that*

$$\bar{B}_{t,(i,j)} = p(y_{t+1:T}|\bar{x}_t = (i,j)) \qquad t = 1, 2, \ldots, T \tag{4.3.19}$$

$$\bar{F}_{t,(i,j)} = p(y_{1:t}, \bar{x}_t = (i,j)) \qquad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, \bar{N}^{(i)}. \tag{4.3.20}$$

Then the HSMM messages for each time $t = 1, 2, \ldots, T$ and each index $i = 1, 2, \ldots, N$ can be computed as

$$F_{t+1,i} = \sum_{j=1}^{\bar{N}^{(i)}} \bar{b}_j^{(i)} \bar{F}_{t,(i,j)} p(y_{t+1}|\theta^{(i)}) \qquad F^*_{t,i} = \sum_{j=1}^{N} F_{t,i} A_{ji} \tag{4.3.21}$$

$$B^*_{t,i} = \sum_{j=1}^{\bar{N}^{(i)}} \bar{c}_j^{(i)} \bar{B}_{t+1,(i,j)} p(y_{t+1}|\theta^{(i)}) \qquad B_{t,i} = \sum_{j=1}^{N} B^*_{t,i} A_{ij}. \tag{4.3.22}$$

Proof. Note that the expressions for $F^*_{t,i}$ and $B_{t,i}$ are identical to those in (4.3.2) and (4.3.4), so it suffices to check the expressions for $B^*_{t,i}$ and $F_{t+1,i}$. Using (4.3.10) and (4.3.11) we have

$$\sum_{j=1}^{N^{(i)}} \bar{b}_j^{(i)} \bar{F}_{t,(i,j)} p(y_{t+1}|\theta^{(i)}) = \sum_{j=1}^{N^{(i)}} p(y_{1:t+1}, \bar{x}_t = (i,j)) p(x_t \neq x_{t+1} | \bar{x}_t = (i,j)) \tag{4.3.23}$$

$$= p(y_{1:t+1}, x_t = i, x_t \neq x_{t+1}) = F_{t+1,i} \tag{4.3.24}$$

$$\sum_{j=1}^{N^{(i)}} \bar{c}_j^{(i)} \bar{B}_{t+1,(i,j)} p(y_{t+1}|\theta^{(i)}) = \sum_{j=1}^{N^{(i)}} p(y_{t+1:T} | \bar{x}_{t+1} = (i,j)) \cdot p(\bar{x}_{t+1} = (i,j) | x_{t+1} = i, x_t \neq x_{t+1}) \tag{4.3.25}$$

$$= p(y_{t+1:T} | x_{t+1} = i, x_t \neq x_{t+1}) = B^*_{t,i} \tag{4.3.26}$$

□

Note that, since the full HMM messages do not need to be stored, the recursions in Eqs. (4.3.21) and (4.3.22) can offer memory savings relative to simply computing the HMM messages in the HMM embedding; instead, the recursions in Eqs. (4.3.21) and (4.3.22) require only $\mathcal{O}(TN + \bar{N})$ memory to compute. In the case of the embedding of Example 4.3.2, since multiplication by each $\bar{A}^{(i)}$ only performs shifting, the HSMM messages can be computed with $\mathcal{O}(TN)$ memory, and indeed in that case the computation corresponds precisely to implementing the recursions in (4.3.1)-(4.3.5). The recursions in Eqs. (4.3.21) and (4.3.22) can be computed in time $\mathcal{O}(TN\bar{N}^2_{\max} + TN^2)$, where $\bar{N}_{\max} = \max_i \bar{N}^{(i)}$.

From the HMM embedding perspective, the HSMM message passing complexity is due to generic duration distributions requiring each HSMM state to be augmented by T pseudostates in the embedding. In the next section we study a particular family of duration distributions for which an HSMM can be encoded as an HMM using many fewer pseudostates.

■ 4.4 HSMM inference with negative binomial durations

In this section, we develop an HMM embedding for HSMMs with negative binomial duration distributions as well as a Gibbs sampler for such models. First, we develop the HMM embedding in terms of the general results of the previous section, prove its correctness, and state its message passing computational complexity. Next, we show how to use the embedding to construct an HSMM Gibbs sampling algorithm in which the overall time complexity of each iteration scales only linearly with the sequence length T. Finally, we show how to generalize the HMM embedding construction to include HSMMs in which the durations are mixtures of negative binomial distributions.

■ 4.4.1 An embedding for negative binomial durations

The negative binomial family of discrete distributions, denoted $\mathrm{NB}(r, p)$ for parameters $r > 0$ and $0 < p < 1$, is well-studied in both frequentist and Bayesian analysis [38]. Furthermore, it has been recommended as a natural model for discrete durations [26, 61] because of it can separately parameterize mean and variance and because it includes geometric durations as a special case when $r = 1$. In this section, using our formulation of HMM embeddings from Section 4.3 we show that negative binomial distributions also provide computational advantages for HSMM message passing and inference.

The negative binomial probability mass function (PMF) can be written[1]

$$p(k|r,p) = \binom{k+r-2}{k-1}(1-p)^r p^{k-1} \qquad k = 1, 2, \ldots. \tag{4.4.1}$$

When r is taken to be fixed, the family of distributions over p is an exponential family:

$$p(k|r,p) = h_r(k) \exp\left\{\eta(p) \cdot t(k) - Z_r(p)\right\} \tag{4.4.2}$$

with

$$h_r(k) \triangleq \binom{k+r-2}{k-1}, \qquad \eta(p) \triangleq \ln p, \tag{4.4.3}$$

$$t(k) \triangleq k - 1, \qquad Z_r(p) \triangleq r \ln(1-p). \tag{4.4.4}$$

However, when considered as a family over (r, p), the negative binomial is no longer an exponential family of distributions because the log base measure $\ln h_r(k)$ has a dependence on r that does not interact linearly with a statistic of the data. The

[1] While some definitions take the support of the negative binomial PMF to be $\{0, 1, 2, \ldots\}$, for duration modeling we shift the PMF to start at 1 because we do not want to include durations of 0 length.

definition of the negative binomial can be generalized to any positive real r parameter by replacing the definition of $h_r(k)$ with the appropriate ratio of gamma functions, but in this chapter we restrict our attention to the case where r is a positive integer. This restriction is essential to the algorithms developed here and does not substantially reduce the negative binomial's expressiveness as a duration model. For a discussion of general results on exponential families of distributions, see Section 2.2.

To construct an efficient HMM embedding, we use the fact that a negative binomial-distributed random variable can be represented as a sum of r geometrically-distributed, shifted random variables:

$$x \sim \mathrm{NB}(r,p) \quad \Longleftrightarrow \quad x = 1 + \sum_{i=1}^{r} z_i \quad \text{with} \quad z_i \overset{\text{iid}}{\sim} \mathrm{ShiftedGeo}(1-p) \tag{4.4.5}$$

where $z_i \overset{\text{iid}}{\sim} \mathrm{ShiftedGeo}(1-p)$ denotes that the z_i are independent and each has a geometric distribution with parameter $1-p$ shifted so that the support includes 0, i.e. the PMF of each z_i is given by

$$p(z|p) = p^z(1-p) \qquad z = 0, 1, 2, \ldots. \tag{4.4.6}$$

Therefore, given an HSMM in which the duration of state i is sampled from $\mathrm{NB}(r^{(i)}, p^{(i)})$, we can construct an HMM embedding by augmenting each HSMM state with $\bar{N}^{(i)} = r^{(i)}$ pseudostates and choosing

$$\bar{A}^{(i)} = \begin{pmatrix} p^{(i)} & 1-p^{(i)} & & \\ & \ddots & \ddots & \\ & & p^{(i)} & 1-p^{(i)} \\ & & & p^{(i)} \end{pmatrix} \qquad \bar{b}^{(i)} = \begin{pmatrix} \\ \\ \\ 1-p^{(i)} \end{pmatrix} \tag{4.4.7}$$

$$\bar{c}^{(i)} = \begin{pmatrix} \mathrm{Binom}(r^{(i)}-1|r^{(i)}-1, p^{(i)}) \\ \mathrm{Binom}(r^{(i)}-2|r^{(i)}-1, p^{(i)}) \\ \vdots \\ \mathrm{Binom}(0|r^{(i)}-1, p^{(i)}) \end{pmatrix} \tag{4.4.8}$$

where $\mathrm{Binom}(k|n,p)$ denotes a binomial PMF with parameters (n,p) and evaluated at k, given by

$$\mathrm{Binom}(k|n,p) \triangleq \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.4.9}$$

In the next proposition, we show that this HMM embedding encodes the correct duration distributions.

Proposition 4.4.1. *Using $\bar{A}^{(i)}$, $\bar{b}^{(i)}$, and $\bar{c}^{(i)}$ as in (4.4.7) and (4.4.8) in the HMM embedding construction of Definition 4.3.1 gives a valid HMM embedding of an HSMM in which the duration of state i is distributed as* $\mathrm{NB}(r^{(i)}, p^{(i)})$.

Proof. By Proposition 4.3.3 it suffices to show that the probability of absorption after exactly d timesteps in each Markov chain with transition matrix and initial state distribution given by

$$\begin{pmatrix} \bar{A}^{(i)} & \bar{b}^{(i)} \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \bar{c}^{(i)\mathsf{T}} & 0 \end{pmatrix}, \tag{4.4.10}$$

respectively, is distributed as $\mathrm{NB}(r^{(i)}, p^{(i)})$. To simplify notation, we drop the superscript i for the remainder of this proof.

Writing X_j to denote the random time to absorption from state $r-j$ for $j = 0, 1, \ldots, r-1$ in the Markov chain parameterized by (4.4.10) , note that by the construction of the transition matrix we can write $X_j = 1 + Y_j + X_{j-1}$ for $j = 1, 2, \ldots, r-1$ and $X_0 = 1 + Y_0$, where $Y_j \overset{\text{iid}}{\sim} \mathrm{Geo}(p)$ for $j = 0, 1, \ldots, r-1$. Therefore $X_j = 1 + j + \sum_{k=1}^{j} Y_j$, and so we can write the PMF of X_j as $\binom{k-1}{k-(r-j)}(1-p)^{r-j}p^{k-(r-j)}$ for $k = r, r+1, \ldots$.

Summing over the initial state distribution we can write probability of absorption after k steps as

$$\sum_{j=0}^{r-1} \binom{r-1}{r-j-1}(1-p)^j p^{r-j-1} \cdot \binom{k-1}{k-(r-j)}(1-p)^{r-j}p^{k-(r-j)} \tag{4.4.11}$$

$$= \sum_{j=0}^{r-1} \binom{r-1}{r-j-1}\binom{k-1}{k-(r-j)}(1-p)^r p^{k-1} \tag{4.4.12}$$

$$= \binom{r+k-2}{k-1}(1-p)^r p^{k-1} \tag{4.4.13}$$

where the last line follows from the Vandermonde identity [4]

$$\binom{m+n}{\ell} = \sum_{j=0}^{\ell} \binom{m}{j}\binom{n}{\ell-j}. \tag{4.4.14}$$

□

Note that we can compute matrix-vector products against each $\bar{A}^{(i)}$ in $\mathcal{O}(r^{(i)})$ time. Therefore with the HMM embedding given in (4.4.7) and (4.4.8) and Proposition 4.3.5, for HSMMs with negative binomial durations we can compute the HSMM messages in time $\mathcal{O}(TN^2 + TNR)$, where $R = \max_i r^{(i)}$. This message passing computation avoids the quadratic dependence on T necessary for generic HSMM messages.

This embedding is not unique; indeed, it can be checked that another valid HMM embedding is given by choosing

$$\bar{A}^{(i)} = \begin{pmatrix} p^{(i)} & (1-p^{(i)})(1-(1-p^{(i)})^{r^{(i)}}) & & \\ & \ddots & \ddots & \\ & & p^{(i)} & (1-p^{(i)})(1-(1-p^{(i)})^2) \\ & & & p^{(i)} \end{pmatrix} \tag{4.4.15}$$

$$\bar{b}^{(i)} = \begin{pmatrix} (1-p^{(i)})^{r^{(i)}} \\ \vdots \\ (1-p^{(i)})^2 \\ 1-p^{(i)} \end{pmatrix} \qquad \bar{c}^{(i)} = \begin{pmatrix} 1 \\ \\ \\ \\ \end{pmatrix} \tag{4.4.16}$$

As we describe in Section 4.5.2, with respect to the HSMM forward message recursions these two alternative HMM embeddings for negative binomial durations are analogous to LTI realizations in observable canonical form and controllable canonical form, respectively.

■ 4.4.2 Gibbs sampling

Here we show how the HMM embedding for HSMMs with negative binomial durations can be used to construct an HSMM (or weak limit HDP-HSMM) Gibbs sampler. Unlike the corresponding Gibbs sampler developed for HSMMs with arbitrary duration distributions in Chapter 3, the time complexity of each iteration of this Gibbs sampler scales only linearly in T, requiring $\mathcal{O}(TN^2 + TNR)$ time for each update instead of $\mathcal{O}(TN^2 + T^2N)$.

We describe the resampling steps for both the HSMM label sequence and the negative binomial parameters; the other sampling updates to the observation parameters, transition matrix, and initial state distribution are performed as in Section 3.4. We place priors of the form $p(r^{(i)}, p^{(i)}) = p(r^{(i)})p(p^{(i)})$ over the negative binomial parameters for each $1, 2, \ldots, N$. In particular, we choose the prior over each $r^{(i)}$ to be a generic distribution with finite support $\{1, 2, \ldots, r_{\max}\}$ and parameter $\nu \in \mathbb{R}_+^N$, writing the PMF as

$$p(r|\nu) \propto \exp\left\{\ln \nu^{\mathsf{T}} \mathbb{1}_r\right\} \qquad r = 1, 2, \ldots, r_{\max} \tag{4.4.17}$$

where $\mathbb{1}_r$ denotes an indicator vector of length N with its rth entry set to 1 and its others set to 0. We place beta priors over each $p^{(i)}$ with parameters (a, b), writing

$$p(p|a,b) = \text{Beta}(a,b) = \exp\left\{(a-1)\ln(p) + (b-1)\ln(1-p) - \ln B(a,b)\right\} \tag{4.4.18}$$

for $p \in (0,1)$, where $B(a,b)$ is the beta function defined by

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt. \tag{4.4.19}$$

Resampling the label sequence $x_{1:T}$. Using the sampled values of each $(r^{(i)}, p^{(i)})$ and the other HSMM parameters, as outlined in the previous section we can use the HMM embedding given in (4.4.7)-(4.4.8) and the message passing recursions given in Proposition 4.3.5 to compute the HSMM backward messages in time $\mathcal{O}(TN^2 + TNR)$. Using the HSMM backward messages and the forward sampling algorithm given in Section 3.4.1, we can then construct a block sample of the label sequence $x_{1:T}$ in time $\mathcal{O}(TN)$. These steps require a total of $\mathcal{O}(TN + NR)$ memory to compute.

Resampling the negative binomial parameters $(r^{(i)}, p^{(i)})$. To derive a sampling update for the negative binomial parameters $(r^{(i)}, p^{(i)})$ for each state $i = 1, 2, \ldots, N$ given the sampled label sequence $x_{1:T}$, we first denote the set of durations of label i in the label sequence by $\{d_k\}_{k=1}^D$, where D is the number of times HSMM state i is entered in the label sequence and we drop the explicit index i from the notation. Then the task is then to sample $(r^{(i)}, p^{(i)})$ conditioned on the $\{d_k\}$ and the hyperparameters a, b, and ν.

Suppressing explicit conditioning on hyperparameters from the notation, we can generate such a sample by first sampling $r^{(i)}|\{d_k\}$ and then sampling $p^{(i)}|r^{(i)}, \{d_k\}$. We can sample $r^{(i)}|\{d_k\}$ according to the PMF

$$p(r|\{d_k\}) \propto \int p(r)p(p)p(\{d_k\}|r,p)\,dp \tag{4.4.20}$$

$$= p(r)\prod_{k=1}^{D}\binom{d_k + r - 2}{d_k - 1} \left(\frac{1}{B(a,b)}\int \exp\left\{(a + \textstyle\sum_{k=1}^D d_k)\ln(p) + (b + rD)\ln(1-p)\right\}\,dp\right) \tag{4.4.21}$$

$$= \nu_r \prod_{k=1}^{D}\binom{d_k + r - 2}{d_k - 1}\left(\frac{B\left(a + \sum_{k=1}^D d_k, b + rD\right)}{B(a,b)}\right) \tag{4.4.22}$$

for $r = 1, 2, \ldots, r_{\max}$, where we have used the fact that, for each fixed r, the Beta prior

on p is conjugate to the negative binomial likelihood. Using each sampled value $\hat{r}^{(i)}$ of $r^{(i)}$, we can then sample $p^{(i)}$ according to

$$p(p|\{d_k\}, r = \hat{r}^{(i)}) = \text{Beta}\left(a + \sum_{i=1}^{D} d_i, b + \hat{r}^{(i)} D\right). \tag{4.4.23}$$

Thus, using Eqs. (4.4.22) and (4.4.23) we can resample the negative binomial parameters $(r^{(i)}, p^{(i)})$ for each HSMM state i, completing the Gibbs sampler.

Note that by resampling the negative binomial parameters $r^{(i)}$ we are effectively performing inference over the size of the HMM embedding given in Section 4.4.1, which we only use to compute the HSMM messages during the resampling step for the label sequence $x_{1:T}$. Thus, unlike the ESHMM approach discussed in Section 4.2, we can learn a much broader class of duration distributions than those corresponding to only fixed $r^{(i)}$ parameters while maintaining efficient message passing.

In Chapter 6, we extend the ideas used to derive this Gibbs sampling algorithm to construct a scalable mean field inference algorithm for which approximate updates can also be computed in time linear in the sequence length T.

■ 4.4.3 HMM embeddings for negative binomial mixtures

Here we show how to generalize the HMM embedding for negative binomial durations to an HMM embedding for durations that are mixtures of negative binomials.

Suppose that the duration distribution for HSMM state i has the PMF

$$p(d|\rho, r, p) = \sum_{j=1}^{K^{(i)}} \rho^{(i,j)} \binom{r^{(i,j)} + d - 2}{d-1} (1 - p^{(i,j)})^{r^{(i,j)}} {p^{(i,j)}}^{d-1} \tag{4.4.24}$$

for mixture weights $\rho^{(i,j)}$ and negative binomial parameters $r^{(i,j)}$ and $p^{(i,j)}$, with $j = 1, 2, \ldots, K^{(i)}$ where $K^{(i)}$ is the number of mixture components. Then we can choose the embedding

$$\bar{A}^{(i)} = \begin{pmatrix} \bar{A}^{(i,1)} & & & \\ & \bar{A}^{(i,2)} & & \\ & & \ddots & \\ & & & \bar{A}^{(i,K^{(i)})} \end{pmatrix} \quad \bar{b}^{(i)} = \begin{pmatrix} \bar{b}^{(i,1)} \\ \bar{b}^{(i,2)} \\ \vdots \\ \bar{b}^{(i,K^{(i)})} \end{pmatrix} \quad \bar{c}^{(i)} = \begin{pmatrix} \bar{c}^{(i,1)} \\ \bar{c}^{(i,2)} \\ \vdots \\ \bar{c}^{(i,K^{(i)})} \end{pmatrix} \tag{4.4.25}$$

where

$$\bar{A}^{(i,j)} = \begin{pmatrix} p^{(i,j)} & 1-p^{(i,j)} & & \\ & \ddots & \ddots & \\ & & p^{(i,j)} & 1-p^{(i,j)} \\ & & & p^{(i,j)} \end{pmatrix} \qquad \bar{b}^{(i,j)} = \begin{pmatrix} \\ \\ \\ 1-p^{(i,j)} \end{pmatrix}$$

$$\bar{c}^{(i,j)} = \rho^{(i,j)} \begin{pmatrix} \text{Binom}(r^{(i,j)}-1 | r^{(i,j)}-1, p^{(i,j)}) \\ \text{Binom}(r^{(i,j)}-2 | r^{(i,j)}-1, p^{(i,j)}) \\ \vdots \\ \text{Binom}(0 | r^{(i,j)}-1, p^{(i,j)}) \end{pmatrix}. \tag{4.4.26}$$

That is, we can construct an embedding for any mixture model by a simple composition of embeddings, where the entrance probabilities $\bar{c}^{(i,j)}$ are weighted by the mixture weights $\rho^{(i,j)}$.

The time complexity for message passing in this HMM embedding is $\mathcal{O}(TNR + TN^2)$ where now $R = \max_i \sum_j r^{(i,j)}$. The memory complexity is only $\mathcal{O}(TN + RN)$ because, as with the other HMM embeddings, we do not need to store the corresponding HMM messages. Gibbs sampling updates can be performed using standard methods for mixture models [11].

■ 4.5 Generalizations via LTI system realization

In this section we extend the notion of HMM embedding of HSMMs developed in Section 4.3 to a notion of LTI realization of HSMMs. The contributions in this section are primarily of theoretical interest, though the framework we develop here may lead to efficient HSMM message passing for a greater variety of duration models or to new approximation schemes. In addition, by making connections to LTI systems and positive realization theory, we show the limits of such methods by giving both an example of a duration distribution that can be represented efficiently as an LTI realization but not an HMM embedding as well as an example of a duration distribution that has no efficient representation of either kind.

The remainder of this section is organized as follows. In Section 4.5.1 we review basic definitions and results from LTI system realization theory. In Section 4.5.2 we develop a definition of LTI realizations of HSMMs and show some basic results, including examples that show the limitations of such an approach to HSMM message passing. We also show that HMM embeddings correspond to (normalized) positive realizations, which are more constrained than general LTI realizations.

■ 4.5.1 LTI systems and realizations

In this section we review some basic definitions and state some results from linear system theory [86] and positive linear system theory [8, 25]. We use these definitions and results to define LTI realizations of HSMMs in Section 4.5.2.

We consider discrete-time, single-input single-output (SISO) linear systems of the form[2]:

$$\begin{aligned} z_{t+1} &= \bar{A} z_t + \bar{b} u_t & z_0 &= 0 \\ w_t &= \bar{c}^{\mathsf{T}} z_t & t &= 0, 1, 2, \ldots \end{aligned} \tag{4.5.1}$$

for input signals u, output signals w, and internal state sequence z. We call the triple $(\bar{A}, \bar{b}, \bar{c})$ the parameters of the LTI system, where A is a $K \times K$ matrix and b and c are vectors of length K for some $0 < K < \infty$. Note that the impulse response of the system can be expressed in terms of the parameters as $\bar{c}^{\mathsf{T}} \bar{A}^{t-1} \bar{b}$ for $t = 1, 2, \ldots$ and the transfer function of the system is $\bar{c}^{\mathsf{T}} (zI - \bar{A})^{-1} \bar{b}$ for $z \in \mathbb{C}$.

Definition 4.5.1 (LTI realization)**.** *Given an impulse response h_t with $h_0 = 0$, we say a system of the form (4.5.1) is an* LTI realization *of h_t if $h_t = \bar{c}^{\mathsf{T}} \bar{A}^{t-1} \bar{b}$ for $t = 1, 2, \ldots$.*

Theorem 4.5.2. *An impulse response h_t has an LTI realization of the form (4.5.1) if and only if the corresponding transfer function $H(z) = \sum_{t=1}^{\infty} h_t z^{-t}$ is a strictly proper rational function of z.*

Definition 4.5.3 (Positive realization [8])**.** *An LTI realization is a* positive realization *if for all nonnegative input signals the output signal and internal state sequence are nonnegative, i.e. if $\forall t\ u_t \geq 0 \implies \forall t\ w_t, z_t \geq 0$, where $z_t \geq 0$ is taken entrywise.*

Theorem 4.5.4 (Theorem 1 [8])**.** *A realization $(\bar{A}, \bar{b}, \bar{c})$ is a positive realization if and only if $\bar{A}, \bar{b}, \bar{c} \geq 0$ entrywise.*

As we make precise in the next subsection, HMM embeddings of HSMMs correspond to positive realizations that are also normalized, in the sense that (4.3.9) is required to be row-stochastic.

Definition 4.5.5 (Canonical realizations)**.** *We say an LTI system of the form (4.5.1) is in* controllable canonical form *if the parameters $(\bar{A}, \bar{b}, \bar{c})$ have the form*

[2] We use z_t to denote the system's internal state and w_t to denote its output to avoid confusion with the HSMM label and observation sequences $x_{1:T}$ and $y_{1:T}$, respectively.

$$\bar{A} = \begin{pmatrix} a_1 & a_2 & \cdots & a_{K-1} & a_K \\ 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \end{pmatrix} \qquad \bar{b} = \begin{pmatrix} 1 \\ \\ \\ \\ \\ \end{pmatrix} \qquad \bar{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{K-1} \\ c_K \end{pmatrix} \tag{4.5.2}$$

and we say it is in observable canonical form *if the parameters have the form*

$$\bar{A} = \begin{pmatrix} a_1 & 1 & & \\ a_2 & & 1 & \\ \vdots & & & \ddots \\ a_{K-1} & & & 1 \\ a_K & & & \end{pmatrix} \qquad \bar{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{K-1} \\ b_K \end{pmatrix} \qquad \bar{c} = \begin{pmatrix} 1 \\ \\ \\ \\ \\ \end{pmatrix}. \tag{4.5.3}$$

We note that it is always possible to write LTI realizations of impulse responses corresponding to strictly proper rational transfer functions in these two canonical forms.

■ 4.5.2 LTI realizations of HSMMs

To motivate the definition of LTI realizations of HSMMs, we start from the definition of HMM embeddings developed in Section 4.3 and show that we can relax the requirement that the embedding parameters $\bar{A}^{(i)}$, $\bar{b}^{(i)}$, and $\bar{c}^{(i)}$ be nonnegative and normalized. In particular, note that using the representation (4.3.7) and Proposition 4.3.5, we can write the HSMM forward messages recursion as

$$F_{t+1,i} = \sum_{j=1}^{\bar{N}^{(i)}} \bar{b}_j^{(i)} \bar{F}_{t,(i,j)} p(y_{t+1}|\theta^{(i)}) \qquad F^*_{t+1,i} = \sum_{j=1}^{N} F_{t+1,i} A_{ji} \tag{4.5.4}$$

$$\bar{F}_{t+1,(i,j)} = \sum_{k=1}^{N^{(i)}} \bar{A}_{kj}^{(i)} \bar{F}_{t,(i,j)} p(y_{t+1}|\theta^{(i)}) + \bar{c}_j^{(i)} F^*_{t+1,i} \qquad \bar{F}_{1,(i,j)} = \pi_i^{(0)} \bar{c}_j^{(i)}. \tag{4.5.5}$$

These recursions can be represented as an interconnection of linear systems as shown in Figure 4.1. The system labeled A simply applies right-multiplication by the HSMM transition matrix A to its input vector. The block labeled by z^{-1} denotes a delay block applied to each of its N inputs. Finally, the systems labeled $D^{(i,t)}$ are the single-input single-output linear *time-varying* systems defined by

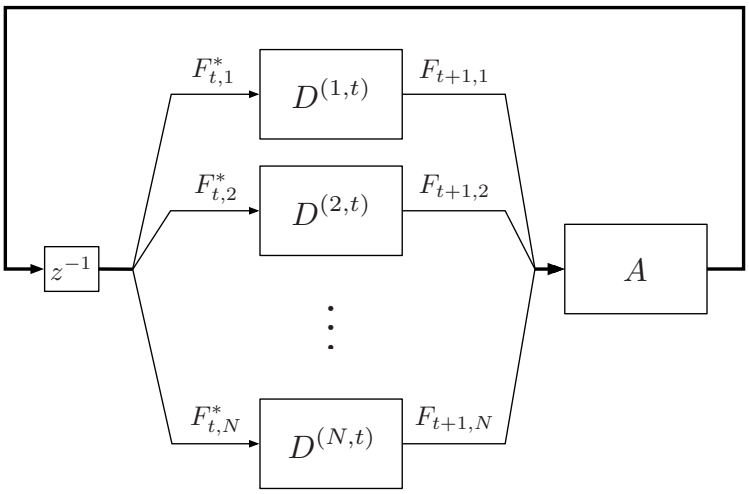

Figure 4.1: A block diagram showing interconnected linear systems corresponding to the forward HSMM message passing recursions in (4.5.4)-(4.5.5). The block labeled z^{-1} denotes a delay block. $F_{t,i}$ and $F^*_{t,i}$ are the standard HSMM forward messages. Note that the inputs and outputs of the blocks labeled z^{-1} and A are vectors of length N, while the blocks labeled $D^{(i,t)}$ operate on each component of the vector.

$$\begin{aligned} z_{t+1} &= \bar{A}^{(i,t)} z_t + \bar{b}^{(i,t)} u_t & z_0 &= 0 \\ w_t &= \bar{c}^{(i)\mathsf{T}} z_t & t &= 0, 1, 2, \ldots. \end{aligned} \tag{4.5.6}$$

for input signal u and output signal w, where

$$\bar{A}^{(i,t)} = p(y_t|\theta^{(i)})\bar{A}^{(i)} \qquad \bar{b}^{(i,t)} = p(y_t|\theta^{(i)})\bar{b}^{(i)}. \tag{4.5.7}$$

The corresponding recursion for the backward messages can be represented similarly.

While the linear systems defined in (4.5.6) are time-varying, as we show in the next lemma to understand when any two such systems yield the same HSMM messages it suffices to compare two corresponding LTI systems.

Lemma 4.5.6. *Let $(\bar{A}, \bar{b}, \bar{c})$ and $(\bar{A}', \bar{b}', \bar{c}')$ be two LTI systems and let $d_t \neq 0$ be any signal that is nonzero everywhere. The LTI systems have the same impulse response if and only if the corresponding time-varying systems of the form*

$$z_{t+1} = d_t(\bar{A} z_t + \bar{b} u_t) \qquad z'_{t+1} = d_t(\bar{A}' z_t + \bar{b}' u_t) \tag{4.5.8}$$

$$w_t = \bar{c}^{\mathsf{T}} z_t \qquad w'_t = \bar{c}'^{\mathsf{T}} z'_t, \tag{4.5.9}$$

with $z_0 = 0$ and $z'_0 = 0$, yield the same output signal $w = w'$ given the same input signal u.

Proof. Because the time-varying systems are linear, we can write [20] the outputs w and w' as linear functions of the input u:

$$w_t = \sum_{\ell=0}^{t-1} d(t,\ell)\bar{c}^{\mathsf{T}}\bar{A}^{t-\ell-1}\bar{b}u_\ell \qquad w'_t = \sum_{\ell=0}^{t-1} d(t,\ell)\bar{c}'^{\mathsf{T}}\bar{A}'^{t-\ell-1}\bar{b}'u_\ell \tag{4.5.10}$$

where $d(t,\ell) = \prod_{k=\ell}^{t} d_k$. If we consider inputs of the form $u_\ell = \delta(k-\ell)$ for some k, where $\delta(k)$ is 1 if k is 0 and 0 otherwise, we see that the outputs w_t and w'_t are equal for all such inputs if and only if the terms in each sum are equal for each ℓ. However, these terms are equal if and only if we have $\bar{c}^{\mathsf{T}}\bar{A}^{t-1}\bar{b} = \bar{c}'^{\mathsf{T}}\bar{A}'^{t-1}\bar{b}'$ for all $t = 1, 2, \ldots$. □

Using Lemma 4.5.6, the following proposition gives a characterization of the parameters $(\bar{A}, \bar{b}, \bar{c})$ that, when used in the recursions (4.5.4)-(4.5.5), compute the correct HSMM messages. This proposition is analogous to Proposition 4.3.3 except it does not require that the realization parameters be entrywise nonnegative or normalized, in the sense that (4.3.9) need not be row-stochastic; it only constrains the system's impulse response and not its parameters or internal state.

Proposition 4.5.7. *Let $(A, \theta, \pi^{(0)}, \vartheta)$ be an HSMM with N states and let $(\bar{A}^{(i)}, \bar{b}^{(i)}, \bar{c}^{(i)})$ be any (not necessarily positive) linear system for each $i = 1, 2, \ldots, N$. If the impulse response of each system satisfies*

$$\bar{c}^{(i)\mathsf{T}}\left(\bar{A}^{(i)}\right)^{d-1}\bar{b}^{(i)} = p(d|\vartheta^{(i)}) \qquad d = 1, 2, \ldots, T-1, \tag{4.5.11}$$

$$\bar{c}^{(i)\mathsf{T}}\left(\bar{A}^{(i)}\right)^{T-1}\bar{b}^{(i)} = p(d \geq T|\vartheta^{(i)}) \tag{4.5.12}$$

then using these systems in the recursions (4.5.4)-(4.5.5) yields the correct HSMM messages.

Proof. If each linear system satisfies (4.5.11)-(4.5.12), then the first T values of the impulse response of each system equal the first T values of the impulse response for the HMM embedding of Example 4.3.2. Since the impulse responses are the same, by Lemma 4.5.6 they yield the same HSMM messages when used in the recursions (4.5.4)-(4.5.5). Therefore the linear systems $(\bar{A}^{(i)}, \bar{b}^{(i)}, \bar{c}^{(i)})$ yield the correct HSMM messages. □

Proposition 4.5.7 motivates the following definition of an LTI realization of an HSMM, which is strictly more general than the definition of HMM embedding be-

cause it allows for parameters $(\bar{A}^{(i)}, \bar{b}^{(i)}, \bar{c}^{(i)})$ that are not necessarily nonnegative and normalized and hence cannot necessarily be interpreted as encoding HMM transition probabilities.

Definition 4.5.8 (LTI realization of an HSMM)**.** *Given an HSMM $(A, \theta, \pi^{(0)}, \vartheta)$ on N states and an observation sequence length T, an* LTI realization *of the HSMM is a set of N LTI systems of the form 4.5.1 with parameters $(\bar{A}^{(i)}, \bar{b}^{(i)}, \bar{c}^{(i)})$ for $i = 1, 2, \ldots, N$ such that*

$$\bar{c}^{(i)\mathsf{T}} \bar{A}^{(i)d-1} \bar{b}^{(i)} = p(d|\vartheta^{(i)}) \qquad d = 1, 2, \ldots, T-1, \tag{4.5.13}$$

$$\bar{c}^{(i)\mathsf{T}} \left(\bar{A}^{(i)}\right)^{T-1} \bar{b}^{(i)} = p(d \geq T|\vartheta^{(i)}). \tag{4.5.14}$$

This definition generalizes the class of representations which can yield efficient HSMM message passing and provides connections to LTI system realization theory. In particular, it makes clear that HMM embeddings correspond to positive realizations of the duration PMF (which are also normalized in the sense that (4.3.9) must also be row-stochastic), for which the internal system state is required to remain nonnegative (and real) for all nonnegative inputs. Definition 4.5.8 removes this requirement of internal nonnegativity, and thus broadens the scope of such efficient representations from positive realizations to general LTI realizations. In particular, the internal state of an LTI system is not required to remain nonnegative or even real-valued.

While this definition is primarily of theoretical interest for the purposes of this chapter, the connection to system realization theory may allow for new algorithms and approximation schemes for HSMM message passing. There are also immediate computational advantages: by showing that the LTI system parameters need not correspond to HMM transition probabilities, it is clear that one can parameterize the system so that each matrix $\bar{A}^{(i)}$ is in bidiagonal Jordan form, and hence the overall HSMM message passing complexity reduces from $\mathcal{O}(TN^2 + TNK_{\max}^2)$ to $\mathcal{O}(TN^2 + TNK_{\max} + K_{\max}^3)$, where $K_{\max}$ denotes the size of the largest matrix $\bar{A}^{(i)}$. This bidiagonalization is not possible with HMM embeddings because HMM transition matrices may have negative and even complex-valued eigenvalues in general.

Definition 4.5.8 also leads to a natural interpretation of the alternative forms of generic HMM embeddings given in Examples 4.3.2 and 4.3.4, as well as the alternative forms of the negative binomial HMM embedding given in Section 4.4. In each of these pairs of alternative embeddings either $\bar{b}^{(i)}$ or $\bar{c}^{(i)}$ is chosen to be an indicator vector, having a single nonzero entry. While these realizations are not precisely in controllable and observable canonical forms because the structure of the corresponding $\bar{A}$ is not in canonical form, the structures of $\bar{b}^{(i)}$ and $\bar{c}^{(i)}$ are analogous to those given in

Definition 4.5.5.

We conclude this section with some examples that use the connection to LTI and positive realization theory to show the limits of both HMM embeddings and LTI realizations for constructing efficient HSMM message passing recursions.

Example 4.5.9. *The Poisson distribution has PMF given by* $p(d|\lambda) = \frac{\lambda^{d-1}}{(d-1)!}e^{-\lambda}$ *for a parameter* $\lambda > 0$ *and* $d = 1, 2, \ldots$. *Its probability generating function (PGF) can be written as* $\sum_{d\geq 1} p(d|\lambda)z^d = e^{-\lambda(1-z)}$. *Since its PGF is irrational, by Theorem 4.5.2 there is no finite LTI realization or HMM embedding of an HSMM with Poisson duration distributions.*[3]

Example 4.5.10. *Adapting Example 4 of Benvenuti and Farina [8], consider a duration distribution with PMF given by*

$$p(d|\psi) = \frac{1}{Z}(1 + \cos[(d-1)\psi])e^{-d} \qquad d \geq 1 \tag{4.5.15}$$

for some $\psi \in \mathbb{R}$ *and a normalization constant* Z. *As shown in Benvenuti and Farina [8], this duration distribution has an LTI realization with parameters*

$$\bar{A} = e^{-1}\begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \bar{b} = \frac{e^{-1}}{Z}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \bar{c} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \tag{4.5.16}$$

but, when ψ/π *is irrational, it has no finite positive realization. Therefore an HSMM with such duration distributions has a finite LTI realization but no finite HMM embedding.*

■ 4.6 Summary

In this chapter we developed a general framework of HMM embeddings for HSMMs and showed how to compute HSMM messages using HMM embeddings. The main practical contribution, which we use in both Chapters 5 and 6, is the construction of an HMM embedding for HSMMs with negative binomial duration distributions. Using this HMM embedding, we showed how to compute HSMM messages in time that scales only linearly with the observation sequence length, and we also derived a complete Gibbs sampler for such HSMMs. The HMM embedding also generalizes to HSMMs with duration distributions that are mixtures of negative binomial distributions.

[3] While there is no finite LTI realization or HMM embedding for all possible sequence lengths T, the generic embedding of Example 4.3.2, in which the number of pseudostates must grow with T and thus message passing complexity is quadratic in T, is always possible.

As a theoretical contribution, we also provided a definition of LTI realizations of HSMMs which is a strict generalization of the notion of HMM embeddings. This generalization may allow for the efficient computation of HSMM messages for a greater variety of duration distributions, and the connections to LTI realization theory may provide a basis for finding efficient approximation algorithms for message passing.