Journal of Computer and System Sciences 78 (2012) 1460–1480

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

A spectral algorithm for learning Hidden Markov Models

Daniel Hsu [a,*], Sham M. Kakade [b], Tong Zhang [a]

[a] Rutgers University, Piscataway, NJ 08854, United States
[b] University of Pennsylvania, Philadelphia, PA 19104, United States

ARTICLE INFO

Article history:
Received 1 February 2010
Received in revised form 31 January 2011
Accepted 22 December 2011
Available online 18 January 2012

Keywords:
Hidden Markov Models
Latent variable models
Observable operator models
Time series
Spectral algorithm
Singular value decomposition
Learning probability distributions
Unsupervised learning

ABSTRACT

Hidden Markov Models (HMMs) are one of the most fundamental and widely used statistical tools for modeling discrete time series. In general, learning HMMs from data is computationally hard (under cryptographic assumptions), and practitioners typically resort to search heuristics which suffer from the usual local optima issues. We prove that under a natural separation condition (bounds on the smallest singular value of the HMM parameters), there is an efficient and provably correct algorithm for learning HMMs. The sample complexity of the algorithm does not explicitly depend on the number of distinct (discrete) observations—it implicitly depends on this quantity through spectral properties of the underlying HMM. This makes the algorithm particularly applicable to settings with a large number of observations, such as those in natural language processing where the space of observation is sometimes the words in a language. The algorithm is also simple, employing only a singular value decomposition and matrix multiplications.

1. Introduction

Hidden Markov Models (HMMs) [2,25] are the workhorse statistical model for discrete time series, with widely diverse applications including automatic speech recognition, natural language processing (NLP), and genomic sequence modeling. In this model, a discrete hidden state evolves according to some Markovian dynamics, and observations at a particular time depend only on the hidden state at that time. The learning problem is to estimate the model only with observation samples from the underlying distribution. Thus far, the predominant learning algorithms have been local search heuristics, such as the Baum–Welch/EM algorithm [3,11].

It is not surprising that practical algorithms have resorted to heuristics, as the general learning problem has been shown to be hard under cryptographic assumptions [30]. Fortunately, the hardness results are for HMMs that seem divorced from those that we are likely to encounter in practical applications.

The situation is in many ways analogous to learning mixture distributions with samples from the underlying distribution. There, the general problem is also believed to be hard. However, much recent progress has been made when certain separation assumptions are made with respect to the component mixture distributions (e.g. [9,10,32,6,4]). Roughly speaking, these separation assumptions imply that with high probability, given a point sampled from the distribution, one can determine the mixture component that generated the point. In fact, there is a prevalent sentiment that we are often only interested in clustering when such a separation condition holds. Much of the theoretical work here has focused on how small this separation can be and still permit an efficient algorithm to recover the model.

We present a simple and efficient algorithm for learning HMMs under a certain natural separation condition. We provide two results for learning. The first is that we can approximate the joint distribution over observation sequences of

* Corresponding author.
E-mail addresses: djhsu@rci.rutgers.edu (D. Hsu), skakade@wharton.upenn.edu (S.M. Kakade), tongz@rci.rutgers.edu (T. Zhang).

0022-0000/$ – see front matter
doi:10.1016/j.jcss.2011.12.025

length t (here, the quality of approximation is measured by total variation distance). As t increases, the approximation quality degrades polynomially. Our second result is on approximating the *conditional* distribution over a future observation, conditioned on some history of observations. We show that this error is asymptotically bounded—i.e. for any t, conditioned on the observations prior to time t, the error in predicting the t-th outcome is controlled. Our algorithm can be thought of as 'improperly' learning an HMM in that we do not explicitly recover the transition and observation models. However, our model does maintain a hidden state representation which is closely (in fact, linearly) related to the HMM's, and can be used for interpreting the hidden state.

The separation condition we require is a spectral condition on both the observation matrix and the transition matrix. Roughly speaking, we require that the observation distributions arising from distinct hidden states be distinct (which we formalize by singular value conditions on the observation matrix). This requirement can be thought of as being weaker than the separation condition for clustering in that the observation distributions can overlap quite a bit—given one observation, we do not necessarily have the information to determine which hidden state it was generated from (unlike in the clustering literature). We also have a spectral condition on the correlation between adjacent observations. We believe both of these conditions to be quite reasonable in many practical applications. Furthermore, given our analysis, extensions to our algorithm which relax these assumptions should be possible.

The algorithm we present has both polynomial sample and computational complexity. Computationally, the algorithm is quite simple—at its core is a singular value decomposition (SVD) of a correlation matrix between past and future observations. This SVD can be viewed as a Canonical Correlation Analysis (CCA) [15] between past and future observations. The sample complexity results we present do not explicitly depend on the number of distinct observations; rather, they implicitly depend on this number through spectral properties of the HMM. This makes the algorithm particularly applicable to settings with a large number of observations, such as those in NLP where the space of observations is sometimes the words in a language.

1.1. Related work

There are two ideas closely related to this work. The first comes from the subspace identification literature in control theory [21,24,18]. The second idea is that, rather than explicitly modeling the hidden states, we can represent the probabilities of sequences of observations as products of matrix observation operators, an idea which dates back to the literature on multiplicity automata [26,5,14].

The subspace identification methods, used in control theory, use spectral approaches to discover the relationship between hidden states and the observations. In this literature, the relationship is discovered for linear dynamical systems such as Kalman filters. The basic idea is that the relationship between observations and hidden states can often be discovered by spectral/SVD methods correlating the past and future observations (in particular, such methods often do a CCA between the past and future observations). However, algorithms presented in the literature cannot be directly used to learn HMMs because they assume additive noise models with noise distributions independent of the underlying states, and such models are not suitable for HMMs (an exception is [1]). In our setting, we use this idea of performing a CCA between past and future observations to uncover information about the observation process (this is done through an SVD on a correlation matrix between past and future observations). The state-independent additive noise condition is avoided through the second idea.

The second idea is that we can represent the probability of sequences as products of matrix operators, as in the literature on multiplicity automata [26,5,14] (see [12] for discussion of this relationship). This idea was re-used in both the Observable Operator Model of Jaeger [16] and the Predictive State Representations of Littman et al. [20], both of which are closely related and both of which can model HMMs. In fact, the former work by Jaeger [16] provides a non-iterative algorithm for learning HMMs, with an asymptotic analysis. However, this algorithm assumed knowing a set of 'characteristic events', which is a rather strong assumption that effectively reveals some relationship between the hidden states and observations. In our algorithm, this problem is avoided through the first idea.

Some of the techniques in the work in [13] for tracking belief states in an HMM are used here. As discussed earlier, we provide a result showing how the model's conditional distributions over observations (conditioned on a history) do not asymptotically diverge. This result was proven in [13] when an approximate model is *already known*. Roughly speaking, the reason this error does not diverge is that the previous observations are always revealing information about the next observation; so with some appropriate contraction property, we would not expect our errors to diverge. Our work borrows from this contraction analysis.

Among recent efforts in various communities [1,31,33,8], the only previous efficient algorithm shown to PAC-learn HMMs in a setting similar to ours is due to [23]. Their algorithm for HMMs is a specialization of a more general method for learning phylogenetic trees from leaf observations. While both this algorithm and ours rely on the same rank condition and compute similar statistics, they differ in two significant regards. First, Mossel and Roch [23] were not concerned with large observation spaces, and thus their algorithm assumes the state and observation spaces to have the same dimension. In addition, Mossel and Roch [23] take the more ambitious approach of learning the observation and transition matrices explicitly, which unfortunately results in a less sample-efficient algorithm that injects noise to artificially spread apart the eigenspectrum of a probability matrix. Our algorithm avoids recovering the observation and transition matrix explicitly,[1] and instead uses subspace identification to learn an alternative representation.

2. Preliminaries

2.1. Hidden Markov Models

The HMM defines a probability distribution over sequences of hidden states (h_t) and observations (x_t). We write the set of hidden states as $[m] = \{1, \ldots, m\}$ and set of observations as $[n] = \{1, \ldots, n\}$, where $m \leqslant n$.

Let $T \in \mathbb{R}^{m\times m}$ be the state transition probability matrix with $T_{ij} = \Pr[h_{t+1} = i | h_t = j]$, $O \in \mathbb{R}^{n\times m}$ be the observation probability matrix with $O_{ij} = \Pr[x_t = i | h_t = j]$, and $\vec{\pi} \in \mathbb{R}^m$ be the initial state distribution with $\vec{\pi}_i = \Pr[h_1 = i]$. The conditional independence properties that an HMM satisfies are: (1) conditioned on the previous hidden state, the current hidden state is sampled independently of all other events in the history; and (2) conditioned on the current hidden state, the current observation is sampled independently from all other events in the history. These conditional independence properties of the HMM imply that T and O fully characterize the probability distribution of any sequence of states and observations.

A useful way of computing the probability of sequences is in terms of 'observation operators', an idea which dates back to the literature on multiplicity automata (see [26,5,14]). The following lemma is straightforward to verify (see [16,13]).

Lemma 1. *For $x = 1, \ldots, n$, define*

$$A_x = T\,\mathrm{diag}(O_{x,1}, \ldots, O_{x,m}).$$

For any t:

$$\Pr[x_1, \ldots, x_t] = \vec{1}_m^\top A_{x_t} \ldots A_{x_1} \vec{\pi}.$$

Our algorithm learns a representation that is based on this observable operator view of HMMs.

2.2. Notation

As already used in Lemma 1, the vector $\vec{1}_m$ is the all-ones vector in $\mathbb{R}^m$. We denote by $x_{1:t}$ the sequence $(x_1, \ldots, x_t)$, and by $x_{t:1}$ its reverse $(x_t, \ldots, x_1)$. When we use a sequence as a subscript, we mean the product of quantities indexed by the sequence elements. So for example, the probability calculation in Lemma 1 can be written $\vec{1}_m^\top A_{x_{t:1}} \vec{\pi}$. We will use $\vec{h}_t$ to denote a probability vector (a distribution over hidden states), with the arrow distinguishing it from the random hidden state variable h_t. Additional notation used in the theorem statements and proofs is listed in Table 1.

2.3. Assumptions

We assume the HMM obeys the following condition.

Condition 1 *(HMM rank condition).* $\vec{\pi} > 0$ element-wise, and O and T are rank m.

The rank condition rules out the problematic case in which some state i has an output distribution equal to a convex combination (mixture) of some other states' output distributions. Such a case could cause a learner to confuse state i with a mixture of these other states. As mentioned before, the general task of learning HMMs (even the specific goal of simply accurately modeling the distribution probabilities [30]) is hard under cryptographic assumptions; the rank condition is a natural way to exclude the malicious instances created by the hardness reduction.

The rank condition on O can be relaxed through a simple modification of our algorithm that looks at multiple observation symbols simultaneously to form the probability estimation tables. For example, if two hidden states have identical observation probability in O but different transition probabilities in T, then they may be differentiated by using two consecutive observations. Although our analysis can be applied in this case with minimal modifications, for clarity, we only state our results for an algorithm that estimates probability tables with rows and columns corresponding to single observations.

2.4. Learning model

Our learning model is similar to those of [19,23] for PAC-learning discrete probability distributions. We assume we can sample observation sequences from an HMM. In particular, we assume each sequence is generated starting from the

[1] In Appendix C, we discuss the key step in [23], and also show how to use their technique in conjunction with our algorithm to recover the HMM observation and transition matrices. Our algorithm does not rely on this extra step—we believe it to be generally unstable—but it can be taken if desired.

same initial state distribution (e.g. the stationary distribution of the Markov chain specified by T). This setting is valid for practical applications including speech recognition, natural language processing, and DNA sequence modeling, where multiple independent sequences are available.

For simplicity, this paper only analyzes an algorithm that uses the initial few observations of each sequence, and ignores the rest. We do this to avoid using concentration bounds with complicated mixing conditions for Markov chains in our sample complexity calculation, as these conditions are not essential to the main ideas we present. In practice, however, one should use the full sequences to form the probability estimation tables required by our algorithm. In such scenarios, a single long sequence is sufficient for learning, and the effective sample size can be simply discounted by the mixing rate of the underlying Markov chain.

Our goal is to derive accurate estimators for the cumulative (joint) distribution $\Pr[x_{1:t}]$ and the conditional distribution $\Pr[x_t|x_{1:t-1}]$ for any sequence length t. For the conditional distribution, we obtain an approximation that does not depend on t, while for the joint distribution, the approximation quality degrades gracefully with t.

3. Observable representations of Hidden Markov Models

A typical strategy for learning HMMs is to estimate the observation and transition probabilities for each hidden state (say, by maximizing the likelihood of a sample). However, since the hidden states are not directly observed by the learner, one often resorts to heuristics (*e.g.* EM) that alternate between imputing the hidden states and selecting parameters $\hat{O}$ and $\hat{T}$ that maximize the likelihood of the sample and current state estimates. Such heuristics can suffer from local optima issues and require careful initialization (e.g. an accurate guess of the hidden states) to avoid failure.

However, under Condition 1, HMMs admit an efficiently learnable parameterization that depends only on *observable quantities*. Because such quantities can be estimated from data, learning this representation avoids any guesswork about the hidden states and thus allows for algorithms with strong guarantees of success.

This parameterization is natural in the context of Observable Operator Models [16], but here we emphasize its connection to subspace identification.

3.1. Definition

Our HMM representation is defined in terms of the following vector and matrix quantities:

$$[P_1]_i = \Pr[x_1 = i],$$

$$[P_{2,1}]_{ij} = \Pr[x_2 = i,\ x_1 = j],$$

$$[P_{3,x,1}]_{ij} = \Pr[x_3 = i,\ x_2 = x,\ x_1 = j] \quad \forall x \in [n],$$

where $P_1 \in \mathbb{R}^n$ is a vector, and $P_{2,1} \in \mathbb{R}^{n\times n}$ and the $P_{3,x,1} \in \mathbb{R}^{n\times n}$ are matrices. These are the marginal probabilities of observation singletons, pairs, and triples.

The representation further depends on a matrix $U \in \mathbb{R}^{n\times m}$ that obeys the following condition.

Condition 2 *(Invertibility condition).* $U^\top O$ is invertible.

In other words, U defines an m-dimensional subspace that preserves the state dynamics—this will become evident in the next few lemmas.

A natural choice for U is given by the 'thin' SVD of $P_{2,1}$, as the next lemma exhibits.

Lemma 2. *Assume $\vec{\pi} > 0$ and that O and T have column rank m. Then $\operatorname{rank}(P_{2,1}) = m$. Moreover, if U is the matrix of left singular vectors of $P_{2,1}$ corresponding to non-zero singular values, then $\operatorname{range}(U) = \operatorname{range}(O)$, so $U \in \mathbb{R}^{n\times m}$ obeys Condition 2.*

Proof. Using the conditional independence properties of the HMM, entries of the matrix $P_{2,1}$ can be factored as

$$[P_{2,1}]_{ij} = \sum_{k=1}^{m}\sum_{\ell=1}^{m} \Pr[x_2 = i,\ x_1 = j,\ h_2 = k,\ h_1 = \ell] = \sum_{k=1}^{m}\sum_{\ell=1}^{m} O_{ik}\, T_{k\ell}\, \vec{\pi}_\ell \left[O^\top\right]_{\ell j}$$

so $P_{2,1} = OT\operatorname{diag}(\vec{\pi})O^\top$ and thus $\operatorname{range}(P_{2,1}) \subseteq \operatorname{range}(O)$. The assumptions on O, T, and $\vec{\pi}$ imply that $T\operatorname{diag}(\vec{\pi})O^\top$ has linearly independent rows and that $P_{2,1}$ has m non-zero singular values. Therefore

$$O = P_{2,1}\left(T\operatorname{diag}(\vec{\pi})O^\top\right)^+$$

(where X^+ denotes the Moore–Penrose pseudo-inverse of a matrix X [29]), which in turn implies $\operatorname{range}(O) \subseteq \operatorname{range}(P_{2,1})$. Thus $\operatorname{rank}(P_{2,1}) = \operatorname{rank}(O) = m$, and also $\operatorname{range}(U) = \operatorname{range}(P_{2,1}) = \operatorname{range}(O)$. $\square$

Our algorithm is motivated by Lemma 2 in that we compute the SVD of an empirical estimate of $P_{2,1}$ to discover a U that satisfies Condition 2. We also note that this choice for U can be thought of as a surrogate for the observation matrix O (see Remark 5).

Now given such a matrix U, we can finally define the observable representation:

$$\begin{aligned}
\vec{b}_1 &= U^\top P_1, \\
\vec{b}_\infty &= \left(P_{2,1}^\top U\right)^+ P_1, \\
B_x &= \left(U^\top P_{3,x,1}\right)\left(U^\top P_{2,1}\right)^+ \quad \forall x \in [n].
\end{aligned}$$

3.2. Basic properties

The following lemma shows that the observable representation, parameterized by $\{\vec{b}_\infty, \vec{b}_1, B_1, \dots, B_n\}$, is sufficient to compute the probabilities of any sequence of observations.

Lemma 3 *(Observable HMM representation). Assume the HMM obeys Condition* 1 *and that* $U \in \mathbb{R}^{n\times m}$ *obeys Condition* 2. *Then*:

1. $\vec{b}_1 = (U^\top O)\vec{\pi}$.
2. $\vec{b}_\infty^\top = \vec{1}_m^\top (U^\top O)^{-1}$.
3. $B_x = (U^\top O) A_x (U^\top O)^{-1}$ $\forall x \in [n]$.
4. $\Pr[x_{1:t}] = \vec{b}_\infty^\top B_{x_{t:1}} \vec{b}_1$ $\forall t \in \mathbb{N}, x_1, \dots, x_t \in [n]$.

In addition to joint probabilities, we can compute conditional probabilities using the observable representation. We do so through (normalized) conditional 'internal states' that depend on a history of observations. We should emphasize that these states are *not* in fact probability distributions over hidden states (though the following lemma shows that they are linearly related). As per Lemma 3, the initial state is

$$\vec{b}_1 = \left(U^\top O\right)\vec{\pi}.$$

Generally, for any $t \geqslant 1$, given observations $x_{1:t-1}$ with $\Pr[x_{1:t-1}] > 0$, we define the internal state as:

$$\vec{b}_t = \vec{b}_t(x_{1:t-1}) = \frac{B_{x_{t-1:1}}\vec{b}_1}{\vec{b}_\infty^\top B_{x_{t-1:1}}\vec{b}_1}.$$

The case $t = 1$ is consistent with the general definition of $\vec{b}_t$ because the denominator is $\vec{b}_\infty^\top \vec{b}_1 = \vec{1}_m^\top (U^\top O)^{-1}(U^\top O)\vec{\pi} = \vec{1}_m^\top \vec{\pi} = 1$. The following result shows how these internal states can be used to compute conditional probabilities $\Pr[x_t = i | x_{1:t-1}]$.

Lemma 4 *(Conditional internal states). Assume the conditions in Lemma* 3. *Then, for any time* t:

1. (*Recursive update of states*) *If* $\Pr[x_{1:t}] > 0$, *then*
$$\vec{b}_{t+1} = \frac{B_{x_t}\vec{b}_t}{\vec{b}_\infty^\top B_{x_t}\vec{b}_t}.$$
2. (*Relation to hidden states*)
$$\vec{b}_t = \left(U^\top O\right)\vec{h}_t(x_{1:t-1}),$$
where $[\vec{h}_t(x_{1:t-1})]_i = \Pr[h_t = i | x_{1:t-1}]$ *is the conditional probability of the hidden state at time t given the observations* $x_{1:t-1}$.
3. (*Conditional observation probabilities*)
$$\Pr[x_t | x_{1:t-1}] = \vec{b}_\infty^\top B_{x_t}\vec{b}_t.$$

Remark 5. If U is the matrix of left singular vectors of $P_{2,1}$ corresponding to non-zero singular values, then U acts much like the observation probability matrix O in the following sense:

$$\begin{array}{ll}
\text{Given a conditional state } \vec{b}_t, & \text{Given a conditional hidden state } \vec{h}_t, \\
\Pr[x_t = i | x_{1:t-1}] = [U\vec{b}_t]_i, & \Pr[x_t = i | x_{1:t-1}] = [O\vec{h}_t]_i.
\end{array}$$

To see this, note that $UU^\top$ is the projection operator to range(U). Since range(U) = range(O) (Lemma 2), we have $UU^\top O = O$, so $U\vec{b}_t = U(U^\top O)\vec{h}_t = O\vec{h}_t$.

3.3. Proofs

Proof of Lemma 3. The first claim is immediate from the fact $P_1 = O\vec{\pi}$. For the second claim, we write P_1 in the following unusual (but easily verified) form:

$$P_1^\top = \vec{1}_m^\top T \operatorname{diag}(\vec{\pi}) O^\top = \vec{1}_m^\top (U^\top O)^{-1} (U^\top O) T \operatorname{diag}(\vec{\pi}) O^\top = \vec{1}_m^\top (U^\top O)^{-1} U^\top P_{2,1}.$$

The matrix $U^\top P_{2,1}$ has linearly independent rows (by the assumptions on $\vec{\pi}$, O, T, and the condition on U), so

$$\vec{b}_\infty^\top = P_1^\top (U^\top P_{2,1})^+ = \vec{1}_m^\top (U^\top O)^{-1} (U^\top P_{2,1}) (U^\top P_{2,1})^+ = \vec{1}_m^\top (U^\top O)^{-1}.$$

To prove the third claim, we first express $P_{3,x,1}$ in terms of A_x:

$$P_{3,x,1} = O A_x T \operatorname{diag}(\vec{\pi}) O^\top = O A_x (U^\top O)^{-1} (U^\top O) T \operatorname{diag}(\vec{\pi}) O^\top = O A_x (U^\top O)^{-1} U^\top P_{2,1}.$$

Again, using the fact that $U^\top P_{2,1}$ has full row rank,

$$B_x = (U^\top P_{3,x,1}) (U^\top P_{2,1})^+ = (U^\top O) A_x (U^\top O)^{-1} (U^\top P_{2,1}) (U^\top P_{2,1})^+ = (U^\top O) A_x (U^\top O)^{-1}.$$

The probability calculation in the fourth claim is now readily seen as a telescoping product that reduces to the product in Lemma 1. □

Proof of Lemma 4. The first claim is a simple induction. The second and third claims are also proved by induction as follows. The base case is clear from Lemma 3 since $\vec{h}_1 = \vec{\pi}$ and $\vec{b}_1 = (U^\top O)\vec{\pi}$, and also $\vec{b}_\infty^\top B_{x_1} \vec{b}_1 = \vec{1}_m^\top A_{x_1} \vec{\pi} = \Pr[x_1]$. For the inductive step,

$$\begin{aligned}
\vec{b}_{t+1} &= \frac{B_{x_t} \vec{b}_t}{\vec{b}_\infty^\top B_{x_t} \vec{b}_t} \\
&= \frac{B_{x_t} (U^\top O) \vec{h}_t}{\Pr[x_t | x_{1:t-1}]} \quad \text{(inductive hypothesis)} \\
&= \frac{(U^\top O) A_{x_t} \vec{h}_t}{\Pr[x_t | x_{1:t-1}]} \quad \text{(Lemma 3)} \\
&= (U^\top O) \frac{\Pr[h_{t+1} = \cdot, x_t | x_{1:t-1}]}{\Pr[x_t | x_{1:t-1}]} \\
&= (U^\top O) \frac{\Pr[h_{t+1} = \cdot | x_{1:t}] \Pr[x_t | x_{1:t-1}]}{\Pr[x_t | x_{1:t-1}]} \\
&= (U^\top O) \vec{h}_{t+1}(x_{1:t})
\end{aligned}$$

and

$$\vec{b}_\infty^\top B_{x_{t+1}} \vec{b}_{t+1} = \vec{1}_m^\top A_{x_{t+1}} \vec{h}_{t+1} = \Pr[x_{t+1} | x_{1:t}]$$

(again, using Lemma 3). □

4. Spectral learning of Hidden Markov Models

4.1. Algorithm

The representation in the previous section suggests the algorithm detailed in Fig. 1, which simply uses random samples to estimate the model parameters. Note that in practice, knowing m is not essential because the method presented here tolerates models that are not exactly HMMs, and the parameter m may be tuned using cross-validation. As we discussed earlier, the requirement for independent samples is only for the convenience of our sample complexity analysis.

The model returned by LEARNHMM(m, N) can be used as follows:

- To predict the probability of a sequence:

$$\widehat{\Pr}[x_1, \ldots, x_t] = \hat{b}_\infty^\top \hat{B}_{x_t} \ldots \hat{B}_{x_1} \hat{b}_1.$$

- Given an observation x_t, the 'internal state' update is:

$$\hat{b}_{t+1} = \frac{\hat{B}_{x_t} \hat{b}_t}{\hat{b}_\infty^\top \hat{B}_{x_t} \hat{b}_t}.$$

Algorithm LearnHMM(m, N):
Inputs: m – number of states, N – sample size
Returns: HMM model parameterized by $\{\hat{b}_1, \hat{b}_\infty, \hat{B}_x \ \forall x \in [n]\}$

1. Independently sample N observation triples (x_1, x_2, x_3) from the HMM to form empirical estimates $\hat{P}_1, \hat{P}_{2,1}, \hat{P}_{3,x,1} \ \forall x \in [n]$ of $P_1, P_{2,1}, P_{3,x,1} \ \forall x \in [n]$.
2. Compute the SVD of $\hat{P}_{2,1}$, and let $\hat{U}$ be the matrix of left singular vectors corresponding to the m largest singular values.
3. Compute model parameters:
   (a) $\hat{b}_1 = \hat{U}^\top \hat{P}_1$,
   (b) $\hat{b}_\infty = (\hat{P}_{2,1}^\top \hat{U})^+ P_1$,
   (c) $\hat{B}_x = \hat{U}^\top \hat{P}_{3,x,1}(\hat{U}^\top \hat{P}_{2,1})^+ \ \forall x \in [n]$.

Fig. 1. HMM learning algorithm.

- To predict the conditional probability of x_t given $x_{1:t-1}$:

$$\widehat{\Pr}[x_t | x_{1:t-1}] = \frac{\hat{b}_\infty^\top \hat{B}_{x_t} \hat{b}_t}{\sum_x \hat{b}_\infty^\top \hat{B}_x \hat{b}_t}.$$

Aside from the random sampling, the running time of the learning algorithm is dominated by the SVD computation of an $n \times n$ matrix. The time required for computing joint probability calculations is $O(tm^2)$ for length t sequences—same as if one used the ordinary HMM parameters (O and T). For conditional probabilities, we require some extra work (proportional to n) to compute the normalization factor. However, our analysis shows that this normalization factor is always close to 1 (see Lemma 13), so it can be safely omitted in many applications.

Note that the algorithm does not explicitly ensure that the predicted probabilities lie in the range $[0, 1]$. This is a dreaded problem that has been faced by other methods for learning and using general operator models [16], and a number of heuristic for coping with the problem have been proposed and may be applicable here (see [17] for some recent developments). We briefly mention that in the case of joint probability prediction, clipping the predictions to the interval $[0, 1]$ can only increase the L_1 accuracy, and that the KL accuracy guarantee explicitly requires the predicted probabilities to be non-zero.

4.2. Main results

We now present our main results. The first result is a guarantee on the accuracy of our joint probability estimates for observation sequences. The second result concerns the accuracy of conditional probability estimates—a much more delicate quantity to bound due to conditioning on unlikely events. We also remark that if the probability distribution is only approximately modeled as an HMM, then our results degrade gracefully based on this approximation quality.

4.2.1. Joint probability accuracy

Let $\sigma_m(M)$ denote the m-th largest singular value of a matrix M. Our sample complexity bound will depend polynomially on $1/\sigma_m(P_{2,1})$ and $1/\sigma_m(O)$.

Also, define

$$\epsilon(k) = \min\left\{\sum_{j \in S} \Pr[x_2 = j]\colon\ S \subseteq [n],\ |S| = n - k\right\}, \tag{1}$$

and let

$$n_0(\varepsilon) = \min\bigl\{k\colon\ \epsilon(k) \leqslant \varepsilon\bigr\}.$$

In other words, $n_0(\varepsilon)$ is the minimum number of observations that account for about $1 - \epsilon$ of the total probability mass. Clearly $n_0(\varepsilon) \leqslant n$, but it can often be much smaller in real applications. For example, in many practical applications, the frequencies of observation symbols observe a power law (called Zipf's law) of the form $f(k) \propto 1/k^s$, where $f(k)$ is the frequency of the k-th most frequently observed symbol. If $s > 1$, then $\epsilon(k) = O(k^{1-s})$, and $n_0(\varepsilon) = O(\varepsilon^{1/(1-s)})$ becomes independent of the number of observations n. This means that for such problems, our analysis below leads to a sample complexity bound for the cumulative distribution $\Pr[x_{1:t}]$ that can be independent of n. This is useful in domains with large n such as natural language processing.

Theorem 6. *There exists a constant $C > 0$ such that the following holds. Pick any $0 < \epsilon, \eta < 1$ and $t \geqslant 1$, and let $\varepsilon_0 = \sigma_m(O)\sigma_m(P_{2,1})\epsilon/(4t\sqrt{m})$. Assume the HMM obeys Condition 1, and*

$$N \geqslant C \cdot \frac{t^2}{\epsilon^2} \cdot \left(\frac{m}{\sigma_m(O)^2 \sigma_m(P_{2,1})^4} + \frac{m \cdot n_0(\varepsilon_0)}{\sigma_m(O)^2 \sigma_m(P_{2,1})^2}\right) \cdot \log\frac{1}{\eta}.$$

With probability at least $1-\eta$*, the model returned by the algorithm* LEARNHMM(m, N) *satisfies*

$$\sum_{x_1,\ldots,x_t} \left|\Pr[x_1,\ldots,x_t] - \widehat{\Pr}[x_1,\ldots,x_t]\right| \leqslant \epsilon.$$

The main challenge in proving Theorem 6 is understanding how the estimation errors accumulate in the algorithm's probability calculation. This would have been less problematic if we had estimates of the usual HMM parameters T and O; the fully observable representation forces us to deal with more cumbersome matrix and vector products.

4.2.2. Conditional probability accuracy

In this section, we analyze the accuracy of our conditional probability predictions $\widehat{\Pr}[x_t|x_1,\ldots,x_{t-1}]$. Intuitively, we might hope that these predictive distributions do not become arbitrarily bad over time (as $t \to \infty$). The reason is that while estimation errors propagate into long-term probability predictions (as evident in Theorem 6), the history of observations constantly provides feedback about the underlying hidden state, and this information is incorporated using Bayes' rule (implicitly via our internal state updates).

This intuition was confirmed by Eyal et al. [13], who showed that if one has an approximate model of T and O for the HMM, then under certain conditions, the conditional prediction does not diverge. This condition is the positivity of the 'value of observation' γ, defined as

$$\gamma = \inf_{\vec{v}:\ \|\vec{v}\|_1=1} \|O\vec{v}\|_1.$$

Note that $\gamma \geqslant \sigma_m(O)/\sqrt{n}$, so it is guaranteed to be positive by Condition 1. However, γ can be much larger than what this crude lower bound suggests.

To interpret this quantity γ, consider any two distributions over hidden states $\vec{h}, \hat{h} \in \mathbb{R}^m$. Then $\|O(\vec{h}-\hat{h})\|_1 \geqslant \gamma\|\vec{h}-\hat{h}\|_1$. Regarding $\vec{h}$ as the true hidden state distribution and $\hat{h}$ as the estimated hidden state distribution, this inequality gives a lower bound on the error of the estimated observation distributions under O. In other words, the observation process, on average, reveal errors in our hidden state estimation. The work of Eyal et al. [13] uses this as a contraction property to show how prediction errors (due to using an approximate model) do not diverge. In our setting, this is more difficult as we do not explicitly estimate O nor do we explicitly maintain distributions over hidden states.

We also need the following assumption, which we discuss further following the theorem statement.

Condition 3 *(Stochasticity condition).* For all observations x and all states i and j, $[A_x]_{ij} \geqslant \alpha > 0$.

Theorem 7. *There exists a constant* $C > 0$ *such that the following holds. Pick any* $0 < \epsilon, \eta < 1$*, and let* $\varepsilon_0 = \sigma_m(O)\sigma_m(P_{2,1})\epsilon/(4\sqrt{m})$*. Assume the HMM obeys Conditions* 1 *and* 3*, and*

$$N \geqslant C \cdot \left(\left(\frac{m}{\epsilon^2\alpha^2} + \frac{(\log(2/\alpha))^4}{\epsilon^4\alpha^2\gamma^4}\right) \cdot \frac{m}{\sigma_m(O)^2\sigma_m(P_{2,1})^4} + \frac{1}{\epsilon^2} \cdot \frac{m \cdot n_0(\varepsilon_0)}{\sigma_m(O)^2\sigma_m(P_{2,1})^2}\right) \cdot \log\frac{1}{\eta}.$$

With probability at least $1-\eta$*, then the model returned by* LEARNHMM(m, N) *satisfies, for any time t,*

$$KL\big(\Pr[x_t|x_1,\ldots,x_{t-1}] \,\big\|\, \widehat{\Pr}[x_t|x_1,\ldots,x_{t-1}]\big) = \mathbb{E}_{x_{1:t}}\left[\ln\frac{\Pr[x_t|x_{1:t-1}]}{\widehat{\Pr}[x_t|x_{1:t-1}]}\right] \leqslant \epsilon.$$

To justify our choice of error measure, note that the problem of bounding the errors of conditional probabilities is complicated by the issue of that, over the long run, we may have to condition on a very low probability event. Thus we need to control the relative accuracy of our predictions. This makes the KL-divergence a natural choice for the error measure. Unfortunately, because our HMM conditions are more naturally interpreted in terms of spectral and normed quantities, we end up switching back and forth between KL and L_1 errors via Pinsker-style inequalities (as in [13]). It is not clear to us if a significantly better guarantee could be obtained with a pure L_1 error analysis (nor is it clear how to do such an analysis).

The analysis in [13] (which assumed that approximations to T and O were provided) dealt with this problem of dividing by zero (during a Bayes' rule update) by explicitly modifying the approximate model so that it *never* assigns the probability of any event to be zero (since if this event occurred, then the conditional probability is no longer defined). In our setting, Condition 3 ensures that true model never assigns the probability of any event to be zero. We can relax this condition somewhat (so that we need not quantify over all observations), though we do not discuss this here.

We should also remark that while our sample complexity bound is significantly larger than in Theorem 6, we are also bounding the more stringent KL-error measure on conditional distributions.

Table 1
Summary of notation.

m, n	Number of states and observations
$n_0(\varepsilon)$	Number of significant observations
O, T, A_x	HMM parameters
P_1, $P_{2,1}$, $P_{3,x,1}$	Marginal probabilities
$\hat{P}_1$, $\hat{P}_{2,1}$, $\hat{P}_{3,x,1}$	Empirical marginal probabilities
ϵ_1, $\epsilon_{2,1}$, $\epsilon_{3,x,1}$	Sampling errors [Section 5.1]
$\hat{U}$	Matrix of m left singular vectors of $\hat{P}_{2,1}$
$\tilde{b}_\infty$, $\tilde{B}_x$, $\tilde{b}_1$	True observable parameters using $\hat{U}$ [Section 5.1]
$\hat{b}_\infty$, $\hat{B}_x$, $\hat{b}_1$	Estimated observable parameters using $\hat{U}$
δ_∞, Δ_x, δ_1	Parameter errors [Section 5.1]
Δ	$\sum_x \Delta_x$ [Section 5.1]
$\sigma_m(M)$	m-th largest singular value of matrix M
$\vec{b}_t$, $\hat{b}_t$	True and estimated states [Section 5.3]
$\vec{h}_t$, $\hat{h}_t$, $\hat{g}_t$	$(\hat{U}^\top O)^{-1}\vec{b}_t$, $(\hat{U}^\top O)^{-1}\hat{b}_t$, $\hat{h}_t/(\vec{1}_m^\top \hat{h}_t)$ [Section 5.3]
$\hat{A}_x$	$(\hat{U}^\top O)^{-1}\hat{B}_x(\hat{U}^\top O)$ [Section 5.3]
γ, α	$\inf\{\|Ov\|_1\colon \|v\|_1 = 1\}$, $\min\{[A_x]_{i,j}\}$

4.2.3. Learning distributions ϵ-close to HMMs

Our L_1 error guarantee for predicting joint probabilities still holds if the sample used to estimate $\hat{P}_1, \hat{P}_{2,1}, \hat{P}_{3,x,1}$ come from a probability distribution $\Pr[\cdot]$ that is merely close to an HMM. Specifically, all we need is that there exists some $t_{\max} \geqslant 3$ and some m state HMM with distribution $\Pr^{\mathrm{HMM}}[\cdot]$ such that:

1. $\Pr^{\mathrm{HMM}}$ satisfies Condition 1 (HMM Rank Condition).
2. For all $t \leqslant t_{\max}$, $\sum_{x_{1:t}} |\Pr[x_{1:t}] - \Pr^{\mathrm{HMM}}[x_{1:t}]| \leqslant \epsilon^{\mathrm{HMM}}(t)$.
3. $\epsilon^{\mathrm{HMM}}(2) \ll \frac{1}{2}\sigma_m(P_{2,1}^{\mathrm{HMM}})$.

The resulting error of our learned model $\widehat{\Pr}$ is

$$\sum_{x_{1:t}} \left|\Pr[x_{1:t}] - \widehat{\Pr}[x_{1:t}]\right| \leqslant \epsilon^{\mathrm{HMM}}(t) + \sum_{x_{1:t}} \left|\Pr^{\mathrm{HMM}}[x_{1:t}] - \widehat{\Pr}[x_{1:t}]\right|$$

for all $t \leqslant t_{\max}$. The second term is now bounded as in Theorem 6, with spectral parameters corresponding to $\Pr^{\mathrm{HMM}}$.

4.3. Subsequent work

Following the initial publication of this work, Siddiqi, Boots, and Gordon have proposed various extensions to the LearnHMM algorithm and its analysis [27]. First, they show that the model parameterization used by our algorithm in fact captures the class of HMMs with rank m transition matrices, which is more general than the class of HMMs with m hidden states. Second, they propose extensions for using longer sequences in the parameter estimation, and also for handling real-valued observations. These extensions prove to be useful in both synthetic experiments and an application to tracking with video data.

A recent work of Song, Boots, Siddiqi, Gordon, and Smola provides a kernelization of our model parameterization in the context of Hilbert space embeddings of (conditional) probability distributions, and extends various aspects of the LearnHMM algorithm and analysis to this setting [28]. This extension is also shown to be advantageous in a number of applications.

5. Proofs

Throughout this section, we assume the HMM obeys Condition 1. Table 1 summarizes the notation that will be used throughout the analysis in this section.

5.1. Estimation errors

Define the following sampling error quantities:

$$\epsilon_1 = \|\hat{P}_1 - P_1\|_2,$$
$$\epsilon_{2,1} = \|\hat{P}_{2,1} - P_{2,1}\|_2,$$
$$\epsilon_{3,x,1} = \|\hat{P}_{3,x,1} - P_{3,x,1}\|_2.$$

The following lemma bounds these errors with high probability as a function of the number of observation samples used to form the estimates.

Lemma 8. *If the algorithm independently samples N observation triples from the HMM, then with probability at least* $1-\eta$:

$$\epsilon_1 \leqslant \sqrt{\frac{1}{N}\ln\frac{3}{\eta}} + \sqrt{\frac{1}{N}},$$

$$\epsilon_{2,1} \leqslant \sqrt{\frac{1}{N}\ln\frac{3}{\eta}} + \sqrt{\frac{1}{N}},$$

$$\max_x \epsilon_{3,x,1} \leqslant \sqrt{\frac{1}{N}\ln\frac{3}{\eta}} + \sqrt{\frac{1}{N}},$$

$$\sum_x \epsilon_{3,x,1} \leqslant \min_k\left(\sqrt{\frac{k}{N}\ln\frac{3}{\eta}} + \sqrt{\frac{k}{N}} + 2\epsilon(k)\right) + \sqrt{\frac{1}{N}\ln\frac{3}{\eta}} + \sqrt{\frac{1}{N}},$$

where $\epsilon(k)$ *is defined in* (1).

Proof. See Appendix A. □

The rest of the analysis estimates how the sampling errors affect the accuracies of the model parameters (which in turn affect the prediction quality). We need some results from matrix perturbation theory, which are given in Appendix B.

Let $U \in \mathbb{R}^{n\times m}$ be matrix of left singular vectors of $P_{2,1}$. The first lemma implies that if $\hat{P}_{2,1}$ is sufficiently close to $P_{2,1}$, i.e. $\epsilon_{2,1}$ is small enough, then the difference between projecting to range$(\hat{U})$ and to range(U) is small. In particular, $\hat{U}^\top O$ will be invertible and be nearly as well-conditioned as $U^\top O$.

Lemma 9. *Suppose* $\epsilon_{2,1} \leqslant \varepsilon \cdot \sigma_m(P_{2,1})$ *for some* $\varepsilon < 1/2$. *Let* $\varepsilon_0 = \epsilon_{2,1}^2/((1-\varepsilon)\sigma_m(P_{2,1}))^2$. *Then*:

1. $\varepsilon_0 < 1$.
2. $\sigma_m(\hat{U}^\top \hat{P}_{2,1}) \geqslant (1-\varepsilon)\sigma_m(P_{2,1})$.
3. $\sigma_m(\hat{U}^\top P_{2,1}) \geqslant \sqrt{1-\varepsilon_0}\,\sigma_m(P_{2,1})$.
4. $\sigma_m(\hat{U}^\top O) \geqslant \sqrt{1-\varepsilon_0}\,\sigma_m(O)$.

Proof. The assumptions imply $\varepsilon_0 < 1$. Since $\sigma_m(\hat{U}^\top \hat{P}_{2,1}) = \sigma_m(\hat{P}_{2,1})$, the second claim is immediate from Corollary 22. Let $U \in \mathbb{R}^{n\times m}$ be the matrix of left singular vectors of $P_{2,1}$. For any $x \in \mathbb{R}^m$, $\|\hat{U}^\top U x\|_2 = \|x\|_2\sqrt{1-\|\hat{U}_\perp^\top U\|_2^2} \geqslant \|x\|_2\sqrt{1-\varepsilon_0}$ by Corollary 22 and the fact $\varepsilon_0 < 1$. The remaining claims follow. □

Now we will argue that the estimated parameters $\hat{b}_\infty, \hat{B}_x, \hat{b}_1$ are close to the following true parameters from the observable representation when $\hat{U}$ is used for U:

$$\tilde{b}_\infty = \left(P_{2,1}^\top \hat{U}\right)^+ P_1 = \left(\hat{U}^\top O\right)^{-\top}\vec{1}_m,$$

$$\tilde{B}_x = \left(\hat{U}^\top P_{3,x,1}\right)\left(\hat{U}^\top P_{2,1}\right)^+ = \left(\hat{U}^\top O\right)A_x\left(\hat{U}^\top O\right)^{-1} \quad \text{for } x = 1,\ldots,n,$$

$$\tilde{b}_1 = \hat{U}^\top P_1.$$

By Lemma 3, as long as $\hat{U}^\top O$ is invertible, these parameters $\tilde{b}_\infty, \tilde{B}_x, \tilde{b}_1$ constitute a valid observable representation for the HMM.

Define the following errors of the estimated parameters:

$$\delta_\infty = \left\|\left(\hat{U}^\top O\right)^\top(\hat{b}_\infty - \tilde{b}_\infty)\right\|_\infty = \left\|\left(\hat{U}^\top O\right)^\top \hat{b}_\infty - \vec{1}_m\right\|_\infty,$$

$$\Delta_x = \left\|\left(\hat{U}^\top O\right)^{-1}(\hat{B}_x - \tilde{B}_x)\left(\hat{U}^\top O\right)\right\|_1 = \left\|\left(\hat{U}^\top O\right)^{-1}\hat{B}_x\left(\hat{U}^\top O\right) - A_x\right\|_1,$$

$$\Delta = \sum_x \Delta_x,$$

$$\delta_1 = \left\|\left(\hat{U}^\top O\right)^{-1}(\hat{b}_1 - \tilde{b}_1)\right\|_1 = \left\|\left(\hat{U}^\top O\right)^{-1}\hat{b}_1 - \vec{\pi}\right\|_1.$$

We can relate these to the sampling errors as follows.

Lemma 10. *Assume $\epsilon_{2,1} \leqslant \sigma_m(P_{2,1})/3$. Then:*

$$\delta_\infty \leqslant 4 \cdot \left(\frac{\epsilon_{2,1}}{\sigma_m(P_{2,1})^2} + \frac{\epsilon_1}{3\sigma_m(P_{2,1})} \right),$$

$$\Delta_x \leqslant \frac{8}{\sqrt{3}} \cdot \frac{\sqrt{m}}{\sigma_m(O)} \cdot \left(\Pr[x_2 = x] \cdot \frac{\epsilon_{2,1}}{\sigma_m(P_{2,1})^2} + \frac{\epsilon_{3,x,1}}{3\sigma_m(P_{2,1})} \right),$$

$$\Delta \leqslant \frac{8}{\sqrt{3}} \cdot \frac{\sqrt{m}}{\sigma_m(O)} \cdot \left(\frac{\epsilon_{2,1}}{\sigma_m(P_{2,1})^2} + \frac{\sum_x \epsilon_{3,x,1}}{3\sigma_m(P_{2,1})} \right),$$

$$\delta_1 \leqslant \frac{2}{\sqrt{3}} \cdot \frac{\sqrt{m}}{\sigma_m(O)} \cdot \epsilon_1.$$

Proof. The assumption on $\epsilon_{2,1}$ guarantees that $\hat{U}^\top O$ is invertible (Lemma 9).

We bound $\delta_\infty = \|(O^\top U)(\hat{b}_\infty - \tilde{b}_\infty)\|_\infty$ by $\|O^\top\|_\infty \|U(\hat{b}_\infty - \tilde{b}_\infty)\|_\infty \leqslant \|\hat{b}_\infty - \tilde{b}_\infty\|_2$. Then:

$$\begin{aligned}
\|\hat{b}_\infty - \tilde{b}_\infty\|_2 &= \big\| \big(\hat{P}_{2,1}^\top \hat{U}\big)^+ \hat{P}_1 - \big(P_{2,1}^\top \hat{U}\big)^+ P_1 \big\|_2 \\
&\leqslant \big\| \big(\big(\hat{P}_{2,1}^\top \hat{U}\big)^+ - \big(P_{2,1}^\top \hat{U}\big)^+\big) \hat{P}_1 \big\|_2 + \big\| \big(P_{2,1}^\top \hat{U}\big)^+ (\hat{P}_1 - P_1) \big\|_2 \\
&\leqslant \big\| \big(\big(\hat{P}_{2,1}^\top \hat{U}\big)^+ - \big(P_{2,1}^\top \hat{U}\big)^+\big) \big\|_2 \|\hat{P}_1\|_1 + \big\| \big(P_{2,1}^\top \hat{U}\big)^+ \big\|_2 \|\hat{P}_1 - P_1\|_2 \\
&\leqslant \frac{1+\sqrt{5}}{2} \cdot \frac{\epsilon_{2,1}}{\min\{\sigma_m(\hat{P}_{2,1}), \sigma_m(P_{2,1}^\top \hat{U})\}^2} + \frac{\epsilon_1}{\sigma_m(P_{2,1}^\top \hat{U})},
\end{aligned}$$

where the last inequality follows from Lemma 23. The bound now follows from Lemma 9.

Next for Δ_x, we bound each term $\|(\hat{U}^\top O)^{-1}(\hat{B}_x - \tilde{B}_x)(\hat{U}^\top O)\|_1$ by $\sqrt{m}\|(\hat{U}^\top O)^{-1}(\hat{B}_x - \tilde{B}_x)\hat{U}^\top\|_2 \|O\|_1 \leqslant \sqrt{m}\|(\hat{U}^\top O)^{-1}\|_2 \|\hat{B}_x - \tilde{B}_x\|_2 \|\hat{U}^\top\|_2 \|O\|_1 = \sqrt{m}\|\hat{B}_x - \tilde{B}_x\|_2 / \sigma_m(\hat{U}^\top O)$. To deal with $\|\hat{B}_x - \tilde{B}_x\|_2$, we use the decomposition

$$\begin{aligned}
\|\hat{B}_x - \tilde{B}_x\|_2 &= \big\| \big(\hat{U}^\top P_{3,x,1}\big)\big(\hat{U}^\top P_{2,1}\big)^+ - \big(\hat{U}^\top \hat{P}_{3,x,1}\big)\big(\hat{U}^\top \hat{P}_{2,1}\big)^+ \big\|_2 \\
&\leqslant \big\| \big(\hat{U}^\top P_{3,x,1}\big)\big(\big(\hat{U}^\top P_{2,1}\big)^+ - \big(\hat{U}^\top \hat{P}_{2,1}\big)^+\big) \big\|_2 + \big\| \hat{U}^\top (P_{3,x,1} - \hat{P}_{3,x,1})\big(\hat{U}^\top P_{2,1}\big)^+ \big\|_2 \\
&\leqslant \|P_{3,x,1}\|_2 \cdot \frac{1+\sqrt{5}}{2} \cdot \frac{\epsilon_{2,1}}{\min\{\sigma_m(\hat{P}_{2,1}), \sigma_m(\hat{U}^\top P_{2,1})\}^2} + \frac{\epsilon_{3,x,1}}{\sigma_m(\hat{U}^\top P_{2,1})} \\
&\leqslant \Pr[x_2 = x] \cdot \frac{1+\sqrt{5}}{2} \cdot \frac{\epsilon_{2,1}}{\min\{\sigma_m(\hat{P}_{2,1}), \sigma_m(\hat{U}^\top P_{2,1})\}^2} + \frac{\epsilon_{3,x,1}}{\sigma_m(\hat{U}^\top P_{2,1})},
\end{aligned}$$

where the second inequality uses Lemma 23, and the final inequality uses the fact $\|P_{3,x,1}\|_2 \leqslant \sqrt{\sum_{i,j}[P_{3,x,1}]_{i,j}^2} \leqslant \sum_{i,j}[P_{3,x,1}]_{i,j} = \Pr[x_2 = x]$. Applying Lemma 9 gives the stated bound on Δ_x and also Δ.

Finally, we bound δ_1 by $\sqrt{m}\|(\hat{U}^\top O)^{-1}\hat{U}^\top\|_2 \|\hat{P}_1 - P_1\|_2 \leqslant \sqrt{m}\epsilon_1/\sigma_m(\hat{U}^\top O)$. Again, the stated bound follows from Lemma 9. □

5.2. Proof of Theorem 6

We need to quantify how estimation errors propagate in the probability calculation. Because the joint probability of a length t sequence is computed by multiplying together t matrices, there is a danger of magnifying the estimation errors exponentially. Fortunately, this is not the case: the following lemma shows that these errors accumulate roughly additively.

Lemma 11. *Assume $\hat{U}^\top O$ is invertible. For any time t:*

$$\sum_{x_{1:t}} \big\| \big(\hat{U}^\top O\big)^{-1} (\hat{B}_{x_{t:1}} \hat{b}_1 - \tilde{B}_{x_{t:1}} \tilde{b}_1) \big\|_1 \leqslant (1+\Delta)^t \delta_1 + (1+\Delta)^t - 1.$$

Proof. By induction on t. The base case, that $\|(\hat{U}^\top O)^{-1}(\hat{b}_1 - \tilde{b}_1)\|_1 \leqslant (1+\Delta)^0 \delta_1 + (1+\Delta)^0 - 1 = \delta_1$ is true by definition. For the inductive step, define unnormalized states $\hat{b}_t = \hat{b}_t(x_{1:t-1}) = \hat{B}_{x_{t-1:1}} \hat{b}_1$ and $\tilde{b}_t = \tilde{b}_t(x_{1:t-1}) = \tilde{B}_{x_{t-1:1}} \tilde{b}_1$. Fix $t > 1$, and assume

$$\sum_{x_{1:t-1}} \big\| \big(\hat{U}^\top O\big)^{-1} (\hat{b}_t - \tilde{b}_t) \big\|_1 \leqslant (1+\Delta)^{t-1} \delta_1 + (1+\Delta)^{t-1} - 1.$$

Then, we can decompose the sum over $x_{1:t}$ as

$$\sum_{x_{1:t}} \big\|\big(\hat{U}^\top O\big)^{-1}(\hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{B}_{x_{t:1}}\tilde{b}_1)\big\|_1 = \sum_{x_{1:t}} \big\|\big(\hat{U}^\top O\big)^{-1}\big((\hat{B}_{x_t} - \tilde{B}_{x_t})\tilde{b}_t + (\hat{B}_{x_t} - \tilde{B}_{x_t})(\hat{b}_t - \tilde{b}_t) + \tilde{B}_{x_t}(\hat{b}_t - \tilde{b}_t)\big)\big\|_1,$$

which, by the triangle inequality, is bounded above by

$$\sum_{x_t}\sum_{x_{1:t-1}} \big\|\big(\hat{U}^\top O\big)^{-1}(\hat{B}_{x_t} - \tilde{B}_{x_t})\big(\hat{U}^\top O\big)\big\|_1 \big\|\big(\hat{U}^\top O\big)^{-1}\tilde{b}_t\big\|_1 \tag{2}$$

$$+ \sum_{x_t}\sum_{x_{1:t-1}} \big\|\big(\hat{U}^\top O\big)^{-1}(\hat{B}_{x_t} - \tilde{B}_{x_t})\big(\hat{U}^\top O\big)\big\|_1 \big\|\big(\hat{U}^\top O\big)^{-1}(\hat{b}_t - \tilde{b}_t)\big\|_1 \tag{3}$$

$$+ \sum_{x_t}\sum_{x_{1:t-1}} \big\|\big(\hat{U}^\top O\big)^{-1}\tilde{B}_t\big(\hat{U}^\top O\big)\big(\hat{U}^\top O\big)^{-1}(\hat{b}_t - \tilde{b}_t)\big\|_1. \tag{4}$$

We deal with each double sum individually. For the sums in (2), we use the fact that $\|(\hat{U}^\top O)^{-1}\tilde{b}_t\|_1 = \Pr[x_{1:t-1}]$, which, when summed over $x_{1:t-1}$, is 1. Thus the entire double sum is bounded by Δ by definition. For (3), we use the inductive hypothesis to bound the inner sum over $\|(\hat{U}^\top O)(\hat{b}_t - \tilde{b}_t)\|_1$; the outer sum scales this bound by Δ (again, by definition). Thus the double sum is bounded by $\Delta((1+\Delta)^{t-1}\delta_1 + (1+\Delta)^{t-1} - 1)$. Finally, for sums in (4), we first replace $(\hat{U}^\top O)^{-1}\tilde{B}_t(\hat{U}^\top O)$ with A_{x_t}. Since A_{x_t} has all non-negative entries, we have that $\|A_{x_t}\vec{v}\|_1 \leqslant \vec{1}_m^\top A_{x_t}|\vec{v}|$ for any vector $\vec{v} \in \mathbb{R}^m$, where $|\vec{v}|$ denotes element-wise absolute value of $\vec{v}$. Now the fact $\vec{1}_m^\top \sum_{x_t} A_{x_t}|\vec{v}| = \vec{1}_m^\top T|\vec{v}| = \vec{1}_m^\top|\vec{v}| = \|\vec{v}\|_1$ and the inductive hypothesis imply the double sum in (4) is bounded by $(1+\Delta)^{t-1}\delta_1 + (1+\Delta)^{t-1} - 1$. Combining these bounds for (2), (3), and (4) completes the induction. □

All that remains is to bound the effect of errors in $\hat{b}_\infty$. Theorem 6 will follow from the following lemma combined with the sampling error bounds of Lemma 8.

Lemma 12. *Assume $\epsilon_{2,1} \leqslant \sigma_m(P_{2,1})/3$. Then for any t,*

$$\sum_{x_{1:t}} \big|\Pr[x_{1:t}] - \widehat{\Pr}[x_{1:t}]\big| \leqslant \delta_\infty + (1+\delta_\infty)\big((1+\Delta)^t\delta_1 + (1+\Delta)^t - 1\big).$$

Proof. By Lemma 9 and the condition on $\epsilon_{2,1}$, we have $\sigma_m(\hat{U}^\top O) > 0$ so $\hat{U}^\top O$ is invertible.

Now we can decompose the L_1 error as follows:

$$\begin{aligned}
\sum_{x_{1:t}} \big|\widehat{\Pr}[x_{1:t}] - \Pr[x_{1:t}]\big| &= \sum_{x_{1:t}} \big|\hat{b}_\infty^\top \hat{B}_{x_{t:1}}\hat{b}_1 - \vec{b}_\infty^\top B_{x_{t:1}}\vec{b}_1\big| \\
&= \sum_{x_{1:t}} \big|\hat{b}_\infty^\top \hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{b}_\infty^\top \tilde{B}_{x_{t:1}}\tilde{b}_1\big| \\
&\leqslant \sum_{x_{1:t}} \big|(\hat{b}_\infty - \tilde{b}_\infty)^\top\big(\hat{U}^\top O\big)\big(\hat{U}^\top O\big)^{-1}\tilde{B}_{x_{t:1}}\tilde{b}_1\big| && (5)\\
&\quad + \sum_{x_{1:t}} \big|(\hat{b}_\infty - \tilde{b}_\infty)^\top\big(\hat{U}^\top O\big)\big(\hat{U}^\top O\big)^{-1}(\hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{B}_{x_{t:1}}\tilde{b}_1)\big| && (6)\\
&\quad + \sum_{x_{1:t}} \big|\tilde{b}_\infty^\top\big(\hat{U}^\top O\big)\big(\hat{U}^\top O\big)^{-1}(\hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{B}_{x_{t:1}}\tilde{b}_1)\big|. && (7)
\end{aligned}$$

The first sum (5) is

$$\begin{aligned}
&\sum_{x_{1:t}} \big|(\hat{b}_\infty - \tilde{b}_\infty)^\top\big(\hat{U}^\top O\big)\big(\hat{U}^\top O\big)^{-1}\tilde{B}_{x_{t:1}}\tilde{b}_1\big| \\
&\quad\leqslant \sum_{x_{1:t}} \big\|\big(\hat{U}^\top O\big)^\top(\hat{b}_\infty - \tilde{b}_\infty)\big\|_\infty \big\|\big(\hat{U}^\top O\big)^{-1}\tilde{B}_{x_{t:1}}\tilde{b}_1\big\|_1 \\
&\quad\leqslant \sum_{x_{1:t}} \delta_\infty \|A_{x_{t:1}}\vec{\pi}\|_1 = \sum_{x_{1:t}} \delta_\infty \Pr[x_{1:t}] = \delta_\infty,
\end{aligned}$$

where the first inequality is Hölder's, and the second uses the bounds in Lemma 10.

The second sum (6) employs Hölder's and Lemma 11:

$$\begin{aligned}\left|(\hat{b}_\infty - \tilde{b}_\infty)^\top(\hat{U}^\top O)(\hat{U}^\top O)^{-1}(\hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{B}_{x_{t:1}}\tilde{b}_1)\right| &\leqslant \left\|(\hat{U}^\top O)^\top(\hat{b}_\infty - \tilde{b}_\infty)\right\|_\infty \left\|(\hat{U}^\top O)^{-1}(\hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{B}_{x_{t:1}}\tilde{b}_1)\right\|_1 \\ &\leqslant \delta_\infty\big((1+\Delta)^t\delta_1 + (1+\Delta)^t - 1\big).\end{aligned}$$

Finally, the third sum (7) uses Lemma 11:

$$\begin{aligned}\sum_{x_{1:t}}\left|\tilde{b}_\infty^\top(\hat{U}^\top O)(\hat{U}^\top O)^{-1}(\hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{B}_{x_{t:1}}\tilde{b}_1)\right| &= \sum_{x_{1:t}}\left|1^\top(\hat{U}^\top O)^{-1}(\hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{B}_{x_{t:1}}\tilde{b}_1)\right| \\ &\leqslant \sum_{x_{1:t}}\left\|(\hat{U}^\top O)^{-1}(\hat{B}_{x_{t:1}}\hat{b}_1 - \tilde{B}_{x_{t:1}}\tilde{b}_1)\right\|_1 \\ &\leqslant (1+\Delta)^t\delta_1 + (1+\Delta)^t - 1.\end{aligned}$$

Combining these gives the desired bound. □

Proof of Theorem 6. By Lemma 8, the specified number of samples N (with a suitable constant C), together with the setting of ε in $n_0(\varepsilon)$, guarantees the following sampling error bounds:

$$\begin{aligned}&\epsilon_1 \leqslant \min\big(0.05\cdot(3/8)\cdot\sigma_m(P_{2,1})\cdot\epsilon,\ 0.05\cdot(\sqrt{3}/2)\cdot\sigma_m(O)\cdot(1/\sqrt{m})\cdot\epsilon\big),\\ &\epsilon_{2,1} \leqslant \min\big(0.05\cdot(1/8)\cdot\sigma_m(P_{2,1})^2\cdot(\epsilon/5),\ 0.01\cdot(\sqrt{3}/8)\cdot\sigma_m(O)\cdot\sigma_m(P_{2,1})^2\cdot\big(1/(t\sqrt{m})\big)\cdot\epsilon\big),\\ &\sum_x \epsilon_{3,x,1} \leqslant 0.39\cdot(3\sqrt{3}/8)\cdot\sigma_m(O)\cdot\sigma_m(P_{2,1})\cdot\big(1/(t\sqrt{m})\big)\cdot\epsilon.\end{aligned}$$

These, in turn, imply the following parameter error bounds, via Lemma 10: $\delta_\infty \leqslant 0.05\epsilon$, $\delta_1 \leqslant 0.05\epsilon$, and $\Delta \leqslant 0.4\epsilon/t$. Finally, Lemma 12 and the fact $(1+a/t)^t \leqslant 1+2a$ for $a \leqslant 1/2$, imply the desired L_1 error bound of ϵ. □

5.3. Proof of Theorem 7

In this subsection, we assume the HMM obeys Condition 3 (in addition to Condition 1).

We introduce the following notation. Let the unnormalized estimated conditional hidden state distributions be

$$\hat{h}_t = (\hat{U}^\top O)^{-1}\hat{b}_t,$$

and its normalized version,

$$\hat{g}_t = \hat{h}_t/(\vec{1}_m^\top \hat{h}_t).$$

Also, let

$$\hat{A}_x = (\hat{U}^\top O)^{-1}\hat{B}_x(\hat{U}^\top O).$$

This notation lets us succinctly compare the updates made by our estimated model to the updates of the true model. Our algorithm never explicitly computes these hidden state distributions $\hat{g}_t$ (as it would require knowledge of the unobserved O). However, under certain conditions (namely Conditions 1 and 3 and some estimation accuracy requirements), these distributions are well-defined and thus we use them for sake of analysis.

The following lemma shows that if the estimated parameters are accurate, then the state updates behave much like the true hidden state updates.

Lemma 13. *For any probability vector $\vec{w} \in \mathbb{R}^m$ and any observation x,*

$$\left|\sum_x \hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\vec{w} - 1\right| \leqslant \delta_\infty + \delta_\infty\Delta + \Delta \quad \textit{and}$$

$$\frac{[\hat{A}_x\vec{w}]_i}{\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\vec{w}} \geqslant \frac{[A_x\vec{w}]_i - \Delta_x}{\vec{1}_m^\top A_x\vec{w} + \delta_\infty + \delta_\infty\Delta_x + \Delta_x} \quad \textit{for all } i = 1,\ldots,m.$$

Moreover, for any non-zero vector $\vec{w} \in \mathbb{R}^m$,

$$\frac{\vec{1}_m^\top\hat{A}_x\vec{w}}{\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\vec{w}} \leqslant \frac{1}{1-\delta_\infty}.$$

Proof. We need to relate the effect of the estimated operator $\hat{A}_x$ to that of the true operator A_x. First assume $\vec{w}$ is a probability vector. Then:

$$\begin{aligned}&\big|\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\vec{w} - \vec{1}_m^\top A_x\vec{w}\big|\\ &\quad= \big|(\hat{b}_\infty - \tilde{b}_\infty)^\top(\hat{U}^\top O)A_x\vec{w} + (\hat{b}_\infty - \tilde{b}_\infty)^\top(\hat{U}^\top O)(\hat{A}_x - A_x)\vec{w} + \tilde{b}_\infty(\hat{U}^\top O)(\hat{A}_x - A_x)\vec{w}\big|\\ &\quad\leqslant \big\|(\hat{b}_\infty - \tilde{b}_\infty)^\top(\hat{U}^\top O)\big\|_\infty\|A_x\vec{w}\|_1 + \big\|(\hat{b}_\infty - \tilde{b}_\infty)^\top(\hat{U}^\top O)\big\|_\infty\big\|(\hat{A}_x - A_x)\big\|_1\|\vec{w}\|_1 + \big\|(\hat{A}_x - A_x)\big\|_1\|\vec{w}\|_1.\end{aligned}$$

Therefore we have

$$\left|\sum_x \hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\vec{w} - 1\right| \leqslant \delta_\infty + \delta_\infty\Delta + \Delta$$

and

$$\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\vec{w} \leqslant \vec{1}_m^\top A_x\vec{w} + \delta_\infty + \delta_\infty\Delta_x + \Delta_x.$$

Combining these inequalities with

$$[\hat{A}_x\vec{w}]_i = [A_x\vec{w}]_i + \big[(\hat{A}_x - A_x)\vec{w}\big]_i \geqslant [A_x\vec{w}]_i - \big\|(\hat{A}_x - A_x)\vec{w}\big\|_1 \geqslant [A_x\vec{w}]_i - \big\|(\hat{A}_x - A_x)\big\|_1\|\vec{w}\|_1 \geqslant [A_x\vec{w}]_i - \Delta_x$$

gives the first claim.

Now drop the assumption that $\vec{w}$ is a probability vector, and assume $\vec{1}_m^\top\hat{A}_x\vec{w} \neq 0$ without loss of generality. Then:

$$\frac{\vec{1}_m^\top\hat{A}_x\vec{w}}{\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\vec{w}} = \frac{\vec{1}_m^\top\hat{A}_x\vec{w}}{\vec{1}_m^\top\hat{A}_x\vec{w} + (\hat{b}_\infty - \tilde{b}_\infty)^\top(\hat{U}^\top O)\hat{A}_x\vec{w}} \leqslant \frac{\|\hat{A}_x\vec{w}\|_1}{\|\hat{A}_x\vec{w}\|_1 - \|(\hat{U}^\top O)^\top(\hat{b}_\infty - \tilde{b}_\infty)\|_\infty\|\hat{A}_x\vec{w}\|_1}$$

which is at most $1/(1-\delta_\infty)$ as claimed. □

A consequence of Lemma 13 is that if the estimated parameters are sufficiently accurate, then the state updates never allow predictions of very small hidden state probabilities.

Corollary 14. *Assume $\delta_\infty \leqslant 1/2$, $\max_x \Delta_x \leqslant \alpha/3$, $\delta_1 \leqslant \alpha/8$, and $\max_x \delta_\infty + \delta_\infty\Delta_x + \Delta_x \leqslant 1/3$. Then $[\hat{g}_t]_i \geqslant \alpha/2$ for all t and i.*

Proof. For $t = 1$, we use Lemma 10 to get $\|\vec{h}_1 - \hat{h}_1\|_1 \leqslant \delta_1 \leqslant 1/2$, so Lemma 17 implies that $\|\vec{h}_1 - \hat{g}_1\|_1 \leqslant 4\delta_1$. Then $[\hat{g}_1]_i \geqslant [\vec{h}_1]_i - |[\vec{h}_1]_i - [\hat{g}_1]_i| \geqslant \alpha - 4\delta_1 \geqslant \alpha/2$ (using Condition 3) as needed. For $t > 1$, Lemma 13 implies

$$\frac{[\hat{A}_x\hat{g}_{t-1}]_i}{\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\hat{g}_{t-1}} \geqslant \frac{[A_x\hat{g}_{t-1}]_i - \Delta_x}{\vec{1}_m^\top A_x\hat{g}_{t-1} + \delta_\infty + \delta_\infty\Delta_x + \Delta_x} \geqslant \frac{\alpha - \alpha/3}{1 + 1/3} \geqslant \frac{\alpha}{2}$$

using Condition 3 in the second-to-last step. □

Lemma 13 and Corollary 14 can now be used to prove the contraction property of the KL-divergence between the true hidden states and the estimated hidden states. The analysis shares ideas from [13], though the added difficulty is due to the fact that the state maintained by our algorithm is not a probability distribution.

Lemma 15. *Let $\varepsilon_0 = \max_x 2\Delta_x/\alpha + (\delta_\infty + \delta_\infty\Delta_x + \Delta_x)/\alpha + 2\delta_\infty$. Assume $\delta_\infty \leqslant 1/2$, $\max_x \Delta_x \leqslant \alpha/3$, and $\max_x \delta_\infty + \delta_\infty\Delta_x + \Delta_x \leqslant 1/3$. For all t, if $\hat{g}_t \in \mathbb{R}^m$ is a probability vector, then*

$$KL(\vec{h}_{t+1}\|\hat{g}_{t+1}) \leqslant KL(\vec{h}_t\|\hat{g}_t) - \frac{\gamma^2}{2(\ln\frac{2}{\alpha})^2}KL(\vec{h}_t\|\hat{g}_t)^2 + \varepsilon_0.$$

Proof. The LHS, written as an expectation over $x_{1:t}$, is

$$KL(\vec{h}_{t+1}\|\hat{g}_{t+1}) = \mathbb{E}_{x_{1:t}}\left[\sum_{i=1}^m [\vec{h}_{t+1}]_i \ln\frac{[\vec{h}_{t+1}]_i}{[\hat{g}_{t+1}]_i}\right].$$

We can bound $\ln(1/[\hat{g}_{t+1}]_i)$ as

$$\begin{aligned}
\ln\frac{1}{[\hat{g}_{t+1}]_i} &= \ln\left(\frac{\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t}\hat{g}_t}{[\hat{A}_{x_t}\hat{g}_t]_i}\cdot\vec{1}_m^\top\hat{h}_{t+1}\right)\\
&= \ln\left(\frac{\vec{1}_m^\top A_{x_t}\hat{g}_t}{[A_{x_t}\hat{g}_t]_i}\cdot\frac{[A_{x_t}\hat{g}_t]_i}{[\hat{A}_{x_t}\hat{g}_t]_i}\cdot\frac{\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t}\hat{g}_t}{\vec{1}_m^\top A_{x_t}\hat{g}_t}\cdot\vec{1}_m^\top\hat{h}_{t+1}\right)\\
&\leqslant \ln\left(\frac{\vec{1}_m^\top A_{x_t}\hat{g}_t}{[A_{x_t}\hat{g}_t]_i}\cdot\frac{[A_{x_t}\hat{g}_t]_i}{[A_{x_t}\hat{g}_t]_i-\Delta_{x_t}}\cdot\frac{\vec{1}_m^\top A_{x_t}\hat{g}_t+\delta_\infty+\delta_\infty\Delta_{x_t}+\Delta_{x_t}}{\vec{1}_m^\top A_{x_t}\hat{g}_t}\cdot(1+2\delta_\infty)\right)\\
&\leqslant \ln\left(\frac{\vec{1}_m^\top A_{x_t}\hat{g}_t}{[A_{x_t}\hat{g}_t]_i}\right)+\frac{2\Delta_{x_t}}{\alpha}+\frac{\delta_\infty+\delta_\infty\Delta_{x_t}+\Delta_{x_t}}{\alpha}+2\delta_\infty\\
&\leqslant \ln\left(\frac{\vec{1}_m^\top A_{x_t}\hat{g}_t}{[A_{x_t}\hat{g}_t]_i}\right)+\varepsilon_0,
\end{aligned}$$

where the first inequality follows from Lemma 13, and the second uses $\ln(1+a)\leqslant a$. Therefore,

$$KL(\vec{h}_{t+1}\,\|\,\hat{g}_{t+1})\leqslant\mathbb{E}_{x_{1:t}}\left[\sum_{i=1}^m[\vec{h}_{t+1}]_i\ln\left([\vec{h}_{t+1}]_i\cdot\frac{\vec{1}_m^\top A_{x_t}\hat{g}_t}{[A_{x_t}\hat{g}_t]_i}\right)\right]+\varepsilon_0. \tag{8}$$

The expectation in (8) is teh KL-divergence between $\Pr[h_t|x_{1:t-1}]$ and the distribution over h_{t+1} that is arrived at by updating $\widehat{\Pr}[h_t|x_{1:t-1}]$ (using Bayes' rule) with $\Pr[h_{t+1}|h_t]$ and $\Pr[x_t|h_t]$. Call this second distribution $\widetilde{\Pr}[h_{t+1}|x_{1:t}]$. The chain rule for KL-divergence states

$$\begin{aligned}
&KL\big(\Pr[h_{t+1}|x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_{t+1}|x_{1:t}]\big)+KL\big(\Pr[h_t|h_{t+1},x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_t|h_{t+1},x_{1:t}]\big)\\
&\quad=KL\big(\Pr[h_t|x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_t|x_{1:t}]\big)+\mathrm{KL}\big(\Pr[h_{t+1}|h_t,x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_{t+1}|h_t,x_{1:t}]\big).
\end{aligned}$$

Thus, using the non-negativity of KL-divergence, we have

$$\begin{aligned}
KL\big(\Pr[h_{t+1}|x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_{t+1}|x_{1:t}]\big)&\leqslant KL\big(\Pr[h_t|x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_t|x_{1:t}]\big)+KL\big(\Pr[h_{t+1}|h_t,x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_{t+1}|h_t,x_{1:t}]\big)\\
&=KL\big(\Pr[h_t|x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_t|x_{1:t}]\big),
\end{aligned}$$

where the equality follows from the fact that $\widetilde{\Pr}[h_{t+1}|h_t,x_{1:t}]=\widetilde{\Pr}[h_{t+1}|h_t]=\Pr[h_{t+1}|h_t]=\Pr[h_{t+1}|h_t,x_{1:t}]$. Furthermore,

$$\Pr[h_t=i|x_{1:t}]=\Pr[h_t=i|x_{1:t-1}]\cdot\frac{\Pr[x_t|h_t=i]}{\sum_{j=1}^m\Pr[x_t|h_t=j]\cdot\Pr[h_t=j|x_{1:t-1}]}$$

and

$$\widetilde{\Pr}[h_t=i|x_{1:t}]=\widehat{\Pr}[h_t=i|x_{1:t-1}]\cdot\frac{\Pr[x_t|h_t=i]}{\sum_{j=1}^m\Pr[x_t|h_t=j]\cdot\widehat{\Pr}[h_t=j|x_{1:t-1}]},$$

so

$$\begin{aligned}
&KL\big(\Pr[h_t|x_{1:t}]\,\big\|\,\widetilde{\Pr}[h_t|x_{1:t}]\big)\\
&\quad=\mathbb{E}_{x_{1:t}}\left[\sum_{i=1}^m\Pr[h_t=i|x_{1:t}]\ln\frac{\Pr[h_t=i|x_{1:t-1}]}{\widehat{\Pr}[h_t=i|x_{1:t-1}]}\right]\\
&\qquad-\mathbb{E}_{x_{1:t}}\left[\sum_{i=1}^m\Pr[h_t=i|x_{1:t}]\ln\frac{\sum_{j=1}^m\Pr[x_t|h_t=j]\cdot\Pr[h_t=j|x_{1:t-1}]}{\sum_{j=1}^m\Pr[x_t|h_t=j]\cdot\widehat{\Pr}[h_t=j|x_{1:t-1}]}\right].
\end{aligned}$$

The first expectation is

$$\begin{aligned}
&\mathbb{E}_{x_{1:t}}\left[\sum_{i=1}^m\Pr[h_t=i|x_{1:t}]\ln\frac{\Pr[h_t=i|x_{1:t-1}]}{\widehat{\Pr}[h_t=i|x_{1:t-1}]}\right]\\
&\quad=\mathbb{E}_{x_{1:t-1}}\left[\sum_{x_t}\Pr[x_t|x_{1:t-1}]\sum_{i=1}^m\Pr[h_t=i|x_{1:t}]\ln\frac{\Pr[h_t=i|x_{1:t-1}]}{\widehat{\Pr}[h_t=i|x_{1:t-1}]}\right]
\end{aligned}$$

$$= \mathbb{E}_{x_{1:t-1}}\left[\sum_{x_t}\sum_{i=1}^{m}\Pr[x_t|h_t=i]\cdot\Pr[h_t=i|x_{1:t-1}]\ln\frac{\Pr[h_t=i|x_{1:t-1}]}{\widehat{\Pr}[h_t=i|x_{1:t-1}]}\right]$$

$$= \mathbb{E}_{x_{1:t-1}}\left[\sum_{x_t}\sum_{i=1}^{m}\Pr[x_t,h_t=i|x_{1:t-1}]\ln\frac{\Pr[h_t=i|x_{1:t-1}]}{\widehat{\Pr}[h_t=i|x_{1:t-1}]}\right]$$

$$= KL(\vec{h}_t\|\hat{g}_t),$$

and the second expectation is

$$\mathbb{E}_{x_{1:t}}\left[\sum_{i=1}^{m}\Pr[h_t=i|x_{1:t}]\ln\frac{\sum_{j=1}^{m}\Pr[x_t|h_t=j]\cdot\Pr[h_t=j|x_{1:t-1}]}{\sum_{j=1}^{m}\Pr[x_t|h_t=j]\cdot\widehat{\Pr}[h_t=j|x_{1:t-1}]}\right]$$

$$= \mathbb{E}_{x_{1:t-1}}\left[\sum_{x_t}\Pr[x_t|x_{1:t-1}]\ln\frac{\sum_{j=1}^{m}\Pr[x_t|h_t=j]\cdot\Pr[h_t=j|x_{1:t-1}]}{\sum_{j=1}^{m}\Pr[x_t|h_t=j]\cdot\widehat{\Pr}[h_t=j|x_{1:t-1}]}\right]$$

$$= KL(O\vec{h}_t\|O\hat{g}_t).$$

Substituting these back into (8), we have

$$KL(\vec{h}_{t+1}\|\hat{g}_{t+1}) \leqslant KL(\vec{h}_t\|\hat{g}_t) - KL(O\vec{h}_t\|O\hat{g}_t) + \varepsilon_0.$$

It remains to bound $KL(O\vec{h}_t\|O\hat{g}_t)$ from above. We use Pinsker's inequality [7], which states that for any distributions $\vec{p}$ and $\vec{q}$,

$$KL(\vec{p}\|\vec{q}) \geqslant \frac{1}{2}\|\vec{p}-\vec{q}\|_1^2,$$

together with the definition of γ, to deduce

$$KL(O\vec{h}_t\|O\hat{g}_t) \geqslant \frac{1}{2}\mathbb{E}_{x_{1:t-1}}\|O\vec{h}_t - O\hat{g}_t\|_1^2 \geqslant \frac{\gamma^2}{2}\mathbb{E}_{x_{1:t-1}}\|\vec{h}_t-\hat{g}_t\|_1^2.$$

Finally, by Jensen's inequality and Lemma 18 (the latter applies because of Corollary 14), we have that

$$\mathbb{E}_{x_{1:t-1}}\|\vec{h}_t-\hat{g}_t\|_1^2 \geqslant \left(\mathbb{E}_{x_{1:t-1}}\|\vec{h}_t-\hat{g}_t\|_1\right)^2 \geqslant \left(\frac{1}{\ln\frac{2}{\alpha}}KL(\vec{h}_t\|\hat{g}_t)\right)^2$$

which gives the required bound. □

Finally, the recurrence from Lemma 15 easily gives the following lemma, which in turn combines with the sampling error bounds of Lemma 8 to give Theorem 7.

Lemma 16. *Let* $\varepsilon_0 = \max_x 2\Delta_x/\alpha + (\delta_\infty + \delta_\infty\Delta_x + \Delta_x)/\alpha + 2\delta_\infty$ *and* $\varepsilon_1 = \max_x(\delta_\infty + \sqrt{m}\delta_\infty\Delta_x + \sqrt{m}\Delta_x)/\alpha$. *Assume* $\delta_\infty \leqslant 1/2$, $\max_x \Delta_x \leqslant \alpha/3$, $\max_x \delta_\infty + \delta_\infty\Delta_x + \Delta_x \leqslant 1/3$, $\delta_1 \leqslant \ln(2/\alpha)/(8\gamma^2)$, $\varepsilon_0 \leqslant \ln(2/\alpha)^2/(4\gamma^2)$, *and* $\varepsilon_1 \leqslant 1/2$. *Then for all* t,

$$KL(\vec{h}_t\|\hat{g}_t) \leqslant \max\left(4\delta_1\log(2/\alpha), \sqrt{\frac{2(\ln\frac{2}{\alpha})^2\varepsilon_0}{\gamma^2}}\right) \quad \textit{and}$$

$$KL\big(\Pr[x_t|x_{1:t-1}]\big\|\widehat{\Pr}[x_t|x_{1:t-1}]\big) \leqslant KL(\vec{h}_t\|\hat{g}_t) + \delta_\infty + \delta_\infty\Delta + \Delta + 2\varepsilon_1.$$

Proof. To prove the bound on $KL(\vec{h}_t\|\hat{g}_t)$, we proceed by induction on t. For the base case, Lemmas 18 (with Corollary 14) and 17 imply $KL(\vec{h}_1\|\hat{g}_1) \leqslant \|\vec{h}_1 - \hat{g}_1\|_1\ln(2/\alpha) \leqslant 4\delta_1\ln(2/\alpha)$ as required. The inductive step follows easily from Lemma 15 and simple calculus: assuming $c_2 \leqslant 1/(4c_1)$, $z - c_1z^2 + c_2$ is non-decreasing in z for all $z \leqslant \sqrt{c_2/c_1}$, so $z' \leqslant z - c_1z^2 + c_2$ and $z \leqslant \sqrt{c_2/c_1}$ together imply that $z' \leqslant \sqrt{c_2/c_1}$. The inductive step uses the above fact with $z = KL(\vec{h}_t\|\hat{g}_t)$, $z' = KL(\vec{h}_{t+1}\|\hat{g}_{t+1})$, $c_1 = \gamma^2/(2(\ln(2/\alpha))^2)$, and $c_2 = \max(\varepsilon_0, c_1(4\delta_1\log(2/\alpha))^2)$.

Now we prove the bound on $KL(\Pr[x_t|x_{1:t-1}]\|\widehat{\Pr}[x_t|x_{1:t-1}])$. First, let $\widehat{\Pr}[x_t,h_t|x_{1:t-1}]$ denote our predicted conditional probability of both the hidden state and observation, i.e. the product of the following two quantities:

$$\widehat{\Pr}[h_t=i|x_{1:t-1}] = [\hat{g}_t]_i \quad \text{and} \quad \widehat{\Pr}[x_t|h_t=i,x_{1:t-1}] = \frac{[\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t}]_i}{\sum_x \hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\hat{g}_t}.$$

Now we can apply the chain rule for KL-divergence

$$\begin{aligned}
KL\big(\Pr[x_t|x_{1:t-1}]\,\big\|\,\widehat{\Pr}[x_t|x_{1:t-1}]\big) &\leqslant KL\big(\Pr[h_t|x_{1:t-1}]\,\big\|\,\widehat{\Pr}[h_t|x_{1:t-1}]\big) + KL\big(\Pr[x_t|h_t,x_{1:t-1}]\,\big\|\,\widehat{\Pr}[x_t|h_t,x_{1:t-1}]\big) \\
&= KL(\vec{h}_t\|\hat{g}_t) + \mathbb{E}_{x_{1:t-1}}\left[\sum_{i=1}^{m}\sum_{x_t}[\vec{h}_t]_i O_{x_t,i}\ln\left(O_{x_t,i}\cdot\frac{\sum_x \hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_x\hat{g}_t}{[\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t}]_i}\right)\right] \\
&\leqslant KL(\vec{h}_t\|\hat{g}_t) + \mathbb{E}_{x_{1:t-1}}\left[\sum_{i=1}^{m}\sum_{x_t}[\vec{h}_t]_i O_{x_t,i}\ln\left(\frac{O_{x_t,i}}{[\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t}]_i}\right)\right] \\
&\quad + \ln(1+\delta_\infty+\delta_\infty\Delta+\Delta),
\end{aligned}$$

where the last inequality uses Lemma 13. It will suffice to show that

$$\frac{O_{x_t,i}}{[\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t}]_i} \leqslant 1+2\varepsilon_1.$$

Note that $O_{x_t,i} = [\tilde{b}_\infty^\top(\hat{U}^\top O)A_{x_t}]_i > \alpha$ by Condition 3. Furthermore, for any i,

$$\begin{aligned}
\big|\big[\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t}\big]_i - O_{x_t,i}\big| &\leqslant \big\|\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t} - \tilde{b}_\infty^\top(\hat{U}^\top O)A_{x_t}\big\|_\infty \\
&\leqslant \big\|(\hat{b}_\infty-\tilde{b}_\infty)(\hat{U}^\top O)\big\|_\infty\|A_{x_t}\|_\infty + \big\|(\hat{b}_\infty-\tilde{b}_\infty)(\hat{U}^\top O)\big\|_\infty\|\hat{A}_{x_t}-A_{x_t}\|_\infty \\
&\quad + \big\|\tilde{b}_\infty(\hat{U}^\top O)\big\|_\infty\|\hat{A}_{x_t}-A_{x_t}\|_\infty \\
&\leqslant \delta_\infty + \sqrt{m}\delta_\infty\Delta_{x_t} + \sqrt{m}\Delta_{x_t}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{O_{x_t,i}}{[\hat{b}_\infty^\top(\hat{U}^\top O)\hat{A}_{x_t}]_i} &\leqslant \frac{O_{x_t,i}}{O_{x_t,i} - (\delta_\infty+\sqrt{m}\delta_\infty\Delta_{x_t}+\sqrt{m}\Delta_{x_t})} \\
&\leqslant \frac{1}{1-(\delta_\infty+\sqrt{m}\delta_\infty\Delta_{x_t}+\sqrt{m}\Delta_{x_t})/\alpha} \\
&\leqslant \frac{1}{1-\varepsilon_1} \leqslant 1+2\varepsilon_1
\end{aligned}$$

as needed. □

Proof of Theorem 7. The proof is mostly the same as that of Theorem 6 with $t=1$, except that Lemma 16 introduces additional error terms. Specifically, we require

$$N \geqslant C\cdot\frac{\ln(2/\alpha)^4}{\epsilon^4\alpha^2\gamma^4}\cdot\frac{m}{\sigma_m(O)^2\sigma_m(P_{2,1})^4} \quad\text{and}\quad N \geqslant C\cdot\frac{m}{\epsilon^2\alpha^2}\cdot\frac{m}{\sigma_m(O)^2\sigma_m(P_{2,1})^4}$$

so that the terms

$$\max\left(4\delta_1\log(2/\alpha),\ \sqrt{\frac{2\ln(2/\alpha)^2\varepsilon_0}{\gamma^2}}\right) \quad\text{and}\quad \varepsilon_1,$$

respectively, are $O(\epsilon)$. The specified number of samples N also suffices to imply the preconditions of Lemma 16. The remaining terms are bounded as in the proof of Theorem 6. □

Lemma 17. *If $\|\vec{a}-\vec{b}\|_1 \leqslant c \leqslant 1/2$ and $\vec{b}$ is a probability vector, then $\|\vec{a}/(\vec{1}^\top\vec{a}) - \vec{b}\|_1 \leqslant 4c$.*

Proof. First, it is easy to check that $1-c \leqslant \vec{1}^\top\vec{a} \leqslant 1+c$. Let $I = \{i\colon \vec{a}_i/(\vec{1}^\top\vec{a}) > \vec{b}_i\}$. Then for $i\in I$, $|\vec{a}_i/(\vec{1}^\top\vec{a}) - \vec{b}_i| = \vec{a}_i/(\vec{1}^\top\vec{a}) - \vec{b}_i \leqslant \vec{a}_i/(1-c) - \vec{b}_i \leqslant (1+2c)\vec{a}_i - \vec{b}_i \leqslant |\vec{a}_i - \vec{b}_i| + 2c\vec{a}_i$. Similarly, for $i\notin I$, $|\vec{b}_i - \vec{a}_i/(\vec{1}^\top\vec{a})| = \vec{b}_i - \vec{a}_i/(\vec{1}^\top\vec{a}) \leqslant \vec{b}_i - \vec{a}_i/(1+c) \leqslant \vec{b}_i - (1-c)\vec{a}_i \leqslant |\vec{b}_i - \vec{a}_i| + c\vec{a}_i$. Therefore $\|\vec{a}/(\vec{1}^\top\vec{a}) - \vec{b}\|_1 \leqslant \|\vec{a}-\vec{b}\|_1 + 2c(\vec{1}^\top\vec{a}) \leqslant c + 2c(1+c) \leqslant 4c$. □

Lemma 18. *Let $\vec{a}$ and $\vec{b}$ be probability vectors. If there exists some $c < 1/2$ such that $\vec{b}_i > c$ for all i, then $KL(\vec{a}\|\vec{b}) \leqslant \|\vec{a}-\vec{b}\|_1\log(1/c)$.*

Proof. See [13], Lemma 3.10. □

Acknowledgments

The authors would like to thank John Langford and Ruslan Salakhutdinov for earlier discussions on using bottleneck methods to learn nonlinear dynamic systems; the linearization of the bottleneck idea was the basis of this paper. We also thank Yishay Mansour for pointing out hardness results for learning HMMs. Finally, we thank Geoff Gordon, Byron Boots, and Sajid Siddiqi for alerting us of an error in a previous version of this paper. This work was completed while DH was an intern at TTI-C in 2008. TZ was partially supported by the following grants: AFOSR-10097389, NSF DMS-1007527, and NSF IIS-1016061.

Appendix A. Sample complexity bound

We will assume independent samples to avoid mixing estimation. Otherwise, one can discount the number of samples by one minus the second eigenvalue of the hidden state transition matrix T.

We are bounding the Frobenius norm of the matrix errors. For simplicity, we unroll the matrices into vectors, and use vector notations.

Let z be a discrete random variable that takes values in $\{1, \ldots, d\}$. We are interested in estimating the vector $\vec{q} = [\Pr(z = j)]_{j=1}^{d}$ from N i.i.d. copies z_i of z ($i = 1, \ldots, N$). Let $\vec{q}_i$ be the vector of zeros expect the z_i-th component being one. Then the empirical estimate of $\vec{q}$ is $\hat{q} = \sum_{i=1}^{N} \vec{q}_i / N$. We are interested in bounding the quantity

$$\|\hat{q} - \vec{q}\|_2^2.$$

The following concentration bound is a simple application of McDiarmid's inequality [22].

Proposition 19. *We have* $\forall \epsilon > 0$:

$$\Pr\bigl(\|\hat{q} - \vec{q}\|_2 \geqslant 1/\sqrt{N} + \epsilon\bigr) \leqslant e^{-N\epsilon^2}.$$

Proof. Consider $\hat{q} = \sum_{i=1}^{N} \vec{q}_i / N$, and let $\hat{p} = \sum_{i=1}^{N} \vec{p}_i / N$, where $\vec{p}_i = \vec{q}_i$ except for $i = k$. Then we have $\|\hat{q} - \vec{q}\|_2 - \|\hat{p} - \vec{q}\|_2 \leqslant \|\hat{q} - \hat{p}\|_2 \leqslant \sqrt{2}/N$. By McDiarmid's inequality, we have

$$\Pr\bigl(\|\hat{q} - \vec{q}\|_2 \geqslant \mathbb{E}\|\hat{q} - \vec{q}\|_2 + \epsilon\bigr) \leqslant e^{-N\epsilon^2}.$$

Note that

$$\begin{aligned}
\mathbb{E}\left\|\sum_{i=1}^{N} \vec{q}_i - N\vec{q}\right\|_2 &\leqslant \left(\mathbb{E}\left\|\sum_{i=1}^{N} \vec{q}_i - N\vec{q}\right\|_2^2\right)^{1/2} = \left(\sum_{i=1}^{N} \mathbb{E}\|\vec{q}_i - \vec{q}\|_2^2\right)^{1/2} \\
&= \left(\sum_{i=1}^{N} \mathbb{E}\bigl[1 - 2\vec{q}_i^{\top}\vec{q} + \|\vec{q}\|_2^2\bigr]\right)^{1/2} = \sqrt{N\bigl(1 - \|\vec{q}\|_2^2\bigr)}.
\end{aligned}$$

This leads to the desired bound. $\square$

Using this bound, we obtain with probability $1 - 3\eta$:

$$\begin{aligned}
&\epsilon_1 \leqslant \sqrt{\ln(1/\eta)/N} + \sqrt{1/N}, \\
&\epsilon_{2,1} \leqslant \sqrt{\ln(1/\eta)/N} + \sqrt{1/N}, \\
&\max_x \epsilon_{3,x,1} \leqslant \sqrt{\sum_x \epsilon_{3,x,1}^2} \leqslant \sqrt{\ln(1/\eta)/N} + \sqrt{1/N}, \\
&\sum_x \epsilon_{3,x,1} \leqslant \sqrt{n}\left(\sum_x \epsilon_{3,x,1}^2\right)^{1/2} \leqslant \sqrt{n\ln(1/\eta)/N} + \sqrt{n/N}.
\end{aligned}$$

If the observation dimensionality n is large and sample size N is small, then the third inequality can be improved by considering a more detailed estimate. Given any k, let $\epsilon(k)$ be sum of elements in the smallest $n - k$ probabilities $\Pr[x_2 = x] = \sum_{i,j} [P_{3,x,1}]_{ij}$ (Eq. (1)). Let S_k be the set of these $n - k$ such x. By Proposition 19, we obtain:

$$\sum_{x \notin S_k} \|\hat{P}_{3,x,1} - P_{3,x,1}\|_F^2 + \left|\sum_{x \in S_k} \sum_{i,j} \bigl([\hat{P}_{3,x,1}]_{ij} - [P_{3,x,1}]_{ij}\bigr)\right|^2 \leqslant \bigl(\sqrt{\ln(1/\eta)/N} + \sqrt{1/N}\bigr)^2.$$

Moreover, by the definition of S_k, we have

$$\begin{aligned}\sum_{x\in S_k}\|\hat{P}_{3,x,1}-P_{3,x,1}\|_F &\leqslant \sum_{x\in S_k}\sum_{i,j}\big|[\hat{P}_{3,x,1}]_{ij}-[P_{3,x,1}]_{ij}\big| \\ &\leqslant \sum_{x\in S_k}\sum_{i,j}\max\big(0,[\hat{P}_{3,x,1}]_{ij}-[P_{3,x,1}]_{ij}\big)+\epsilon(k) \\ &\quad+\sum_{x\in S_k}\sum_{i,j}\min\big(0,[\hat{P}_{3,x,1}]_{ij}-[P_{3,x,1}]_{ij}\big)+\epsilon(k) \\ &\leqslant \left|\sum_{x\in S_k}\sum_{i,j}\big([\hat{P}_{3,x,1}]_{ij}-[P_{3,x,1}]_{ij}\big)\right|+2\epsilon(k).\end{aligned}$$

Therefore

$$\sum_x \epsilon_{3,x,1}\leqslant \min_k\big(\sqrt{k\ln(1/\eta)/N}+\sqrt{k/N}+\sqrt{\ln(1/\eta)/N}+\sqrt{1/N}+2\epsilon(k)\big).$$

This means $\sum_x \epsilon_{3,x,1}$ may be small even if n is large, but the number of frequently occurring symbols are small.

Appendix B. Matrix perturbation theory

The following perturbation bounds can be found in [29].

Lemma 20. *(See Theorem 4.11, p. 204 in [29].) Let $A\in\mathbb{R}^{m\times n}$ with $m\geqslant n$, and let $\tilde{A}=A+E$. If the singular values of A and $\tilde{A}$ are $(\sigma_1\geqslant\cdots\geqslant\sigma_n)$ and $(\tilde{\sigma}_1\geqslant\cdots\geqslant\tilde{\sigma}_n)$, respectively, then*

$$|\tilde{\sigma}_i-\sigma_i|\leqslant\|E\|_2,\quad i=1,\ldots,n.$$

Lemma 21. *(See Theorem 4.4, p. 262 in [29].) Let $A\in\mathbb{R}^{m\times n}$, with $m\geqslant n$, with the singular value decomposition $(U_1,U_2,U_3,\Sigma_1,\Sigma_2,V_1,V_2)$:*

$$\begin{bmatrix}U_1^\top\\U_2^\top\\U_3^\top\end{bmatrix}A\,[\,V_1\quad V_2\,]=\begin{bmatrix}\Sigma_1&0\\0&\Sigma_2\\0&0\end{bmatrix}.$$

Let $\tilde{A}=A+E$, with analogous SVD $(\tilde{U}_1,\tilde{U}_2,\tilde{U}_3,\tilde{\Sigma}_1,\tilde{\Sigma}_2,\tilde{V}_1\tilde{V}_2)$. Let Φ be the matrix of canonical angles between $\operatorname{range}(U_1)$ and $\operatorname{range}(\tilde{U}_1)$, and Θ be the matrix of canonical angles between $\operatorname{range}(V_1)$ and $\operatorname{range}(\tilde{V}_1)$. If there exists $\delta,\alpha>0$ such that $\min\sigma(\tilde{\Sigma}_1)\geqslant\alpha+\delta$ and $\max\sigma(\Sigma_2)\leqslant\alpha$, then

$$\max\big\{\|\sin\Phi\|_2,\|\sin\Theta\|_2\big\}\leqslant\frac{\|E\|_2}{\delta}.$$

Corollary 22. *Let $A\in\mathbb{R}^{m\times n}$, with $m\geqslant n$, have rank n, and let $U\in\mathbb{R}^{m\times n}$ be the matrix of n left singular vectors corresponding to the non-zero singular values $\sigma_1\geqslant\cdots\geqslant\sigma_n>0$ of A. Let $\tilde{A}=A+E$. Let $\tilde{U}\in\mathbb{R}^{m\times n}$ be the matrix of n left singular vectors corresponding to the largest n singular values $\tilde{\sigma}_1\geqslant\cdots\geqslant\tilde{\sigma}_n$ of $\tilde{A}$, and let $\tilde{U}_\perp\in\mathbb{R}^{m\times(m-n)}$ be the remaining left singular vectors. Assume $\|E\|_2\leqslant\epsilon\sigma_n$ for some $\epsilon<1$. Then:*

1. $\tilde{\sigma}_n\geqslant(1-\epsilon)\sigma_n$,
2. $\|\tilde{U}_\perp^\top U\|_2\leqslant\|E\|_2/\tilde{\sigma}_n$.

Proof. The first claim follows from Lemma 20, and the second follows from Lemma 21 because the singular values of $\tilde{U}_\perp^\top U$ are the sines of the canonical angles between $\operatorname{range}(U)$ and $\operatorname{range}(\tilde{U})$. □

Lemma 23. *(See Theorem 3.8, p. 143 in [29].) Let $A\in\mathbb{R}^{m\times n}$, with $m\geqslant n$, and let $\tilde{A}=A+E$. Then*

$$\|\tilde{A}^+-A^+\|_2\leqslant\frac{1+\sqrt{5}}{2}\cdot\max\big\{\|A^+\|_2^2,\|\tilde{A}^+\|_2^2\big\}\|E\|_2.$$

Appendix C. Recovering the observation and transition matrices

We sketch how to use the technique of [23] to recover the observation and transition matrices explicitly. This is an extra step that can be used in conjunction with our algorithm.

Define the $n \times n$ matrix $[P_{3,1}]_{i,j} = \Pr[x_3 = i,\ x_1 = j]$. Let $O_x = \operatorname{diag}(O_{x,1}, \ldots, O_{x,m})$, so $A_x = TO_x$. Since $P_{3,x,1} = OA_xT\operatorname{diag}(\vec{\pi})O^\top$, we have $P_{3,1} = \sum_x P_{3,x,1} = OTT\operatorname{diag}(\vec{\pi})O^\top$. Therefore

$$\begin{aligned} U^\top P_{3,x,1} &= U^\top OTO_xT\operatorname{diag}(\vec{\pi})O^\top = (U^\top OT)O_x(U^\top OT)^{-1}(U^\top OT)T\operatorname{diag}(\vec{\pi})O^\top \\ &= (U^\top OT)O_x(U^\top OT)^{-1}(U^\top P_{3,1}). \end{aligned}$$

The matrix $U^\top P_{3,1}$ has full row rank, so $(U^\top P_{3,1})(U^\top P_{3,1})^+ = I$, and thus

$$(U^\top P_{3,x,1})(U^\top P_{3,1})^+ = (U^\top OT)O_x(U^\top OT)^{-1}.$$

Since O_x is diagonal, the eigenvalues of $(U^\top P_{3,x,1})(U^\top P_{3,1})^+$ are exactly the observation probabilities $O_{r,1}, \ldots, O_{r,m}$.

Define i.i.d. random variables $g_x \sim N(0,1)$ for each x. It is shown in [23] that the eigenvalues of

$$\sum_x g_x(U^\top P_{3,x,1})(U^\top P_{3,1})^+ = (U^\top OT)\left(\sum_x g_xO_x\right)(U^\top OT)^{-1}$$

will be separated with high probability (though the separation is roughly on the same order as the failure probability; this is the main source of instability with this method). Therefore an eigen-decomposition will recover the columns of $(U^\top OT)$ up to a diagonal scaling matrix S, *i.e.* $U^\top OTS$. Then for each x, we can diagonalize $(U^\top P_{3,x,1})(U^\top P_{3,1})^+$:

$$(U^\top OTS)^{-1}(U^\top P_{3,x,1})(U^\top P_{3,1})^+(U^\top OTS) = O_x.$$

Now we can form O from the diagonals of O_x. Since O has full column rank, $O^+O = I_m$, so it is now easy to also recover $\vec{\pi}$ and T from P_1 and $P_{2,1}$:

$$O^+P_1 = O^+O\vec{\pi} = \vec{\pi}$$

and

$$O^+P_{2,1}(O^+)^\top\operatorname{diag}(\vec{\pi})^{-1} = O^+(OT\operatorname{diag}(\vec{\pi})O^\top)(O^+)^\top\operatorname{diag}(\vec{\pi})^{-1} = T.$$

Note that because [23] do not allow more observations than states, they do not need to work in a lower dimensional subspace such as range(U). Thus, they perform an eigen-decomposition of the matrix

$$\sum_x g_xP_{3,x,1}P_{3,1}^{-1} = (OT)\left(\sum_x g_xO_x\right)(OT)^{-1},$$

and then use the eigenvectors to form the matrix OT. Thus they rely on the stability of the eigenvectors, which depends heavily on the spacing of the eigenvalues. Consequently, the resulting sample complexity of the algorithm is polynomial in $1/\eta$ (as opposed to $\log(1/\eta)$) where η is the allowed probability of failure.

References

[1] S. Andersson, T. Ryden, R. Johansson, Linear optimal prediction and innovations representations of hidden Markov models, Stochastic Process. Appl. 108 (2003) 131–149.
[2] Leonard E. Baum, J.A. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, Bull. Amer. Math. Soc. 73 (3) (1967) 360–363.
[3] Leonard E. Baum, Ted Petrie, George Soules, Norman Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. of Math. Statist. 41 (1) (1970) 164–171.
[4] S. Charles Brubaker, Santosh Vempala, Isotropic PCA and affine-invariant clustering, in: FOCS, 2008.
[5] J.W. Carlyle, A. Paz, Realization by stochastic finite automaton, J. Comput. System Sci. 5 (1971) 26–40.
[6] Kamalika Chaudhuri, Satish Rao, Learning mixtures of product distributions using correlations and independence, in: COLT, 2008.
[7] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, 1991.
[8] G. Cybenko, V. Crespi, Learning Hidden Markov Models using non-negative matrix factorization, Technical report, 2008, arXiv:0809.4086.
[9] Sanjoy Dasgupta, Learning mixutres of Gaussians, in: FOCS, 1999.
[10] Sanjoy Dasgupta, Leonard Schulman, A probabilistic analysis of EM for mixtures of separated, spherical Gaussians, J. Mach. Learn. Res. 8 (February) (2007) 203–226.
[11] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1) (1977) 1–38.
[12] Even-Dar Eyal, Sham M. Kakade, Yishay Mansour, Planning in POMDPs using multiplicity automata, in: UAI, 2005.
[13] Even-Dar Eyal, Sham M. Kakade, Yishay Mansour, The value of observation for monitoring dynamic systems, in: IJCAI, 2007.
[14] M. Fliess, Matrices deHankel, J. Math. Pures Appl. 53 (1974) 197–222.
[15] H. Hotelling, The most predictable criterion, J. Educ. Psychol. (1935).

[16] Herbert Jaeger, Observable operator models for discrete stochastic time series, Neural Comput. 12 (6) (2000).
[17] H. Jaeger, M. Zhao, K. Kretzschmar, T. Oberstein, D. Popovici, A. Kolling, Learning observable operator models via the es algorithm, in: S. Haykin, J. Principe, T. Sejnowski, J. McWhirter (Eds.), New Directions in Statistical Signal Processing: From Systems to Brain, MIT Press, 2006, pp. 417–464.
[18] T. Katayama, Subspace Methods for System Identification, Springer, 2005.
[19] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, L. Sellie, On the learnability of discrete distributions, in: STOC, 1994, pp. 273–282.
[20] Michael Littman, Richard Sutton, Singh Satinder, Predictive representations of state, in: Advances in Neural Information Processing Systems 14 (NIPS), 2001, pp. 1555–1561.
[21] L. Ljung, System Identification: Theory for the User, Prentice Hall, Englewood Cliffs, NJ, 1987.
[22] C. McDiarmid, On the method of bounded differences, in: Surveys in Combinatorics, Cambridge University Press, 1989, pp. 148–188.
[23] E. Mossel, S. Roch, Learning nonsingular phylogenies and Hidden Markov Models, Ann. Appl. Probab. 16 (2) (2006) 583–614.
[24] P.V. Overschee, B. De Moor, Subspace Identification of Linear Systems, Kluwer Academic Publishers, 1996.
[25] Lawrence R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[26] M.P. Schützenberger, On the definition of a family of automata, Inform. Control 4 (1961) 245–270.
[27] S.M. Siddiqi, B. Boots, G.J. Gordon, Reduced-rank Hidden Markov Models, in: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 2010.
[28] Le Song, Byron Boots, Sajid M. Siddiqi, Geoffrey J. Gordon, Alex Smola, Hilbert space embeddings of Hidden Markov Models, in: Proceedings of the 27th International Conference on Machine Learning, 2010.
[29] G.W. Stewart, Ji-Guang Sun, Matrix Perturbation Theory, Academic Press, New York, 1990.
[30] Sebastiaan Terwijn, On the learnability of Hidden Markov Models, in: International Colloquium on Grammatical Inference, 2002.
[31] B. Vanluyten, J. Willems, B. De Moor, A new approach for the identification of Hidden Markov Models, in: Conference on Decision and Control, 2007.
[32] Santosh Vempala, Grant Wang, A spectral algorithm for learning mixtures of distributions, in: FOCS, 2002.
[33] MingJie Zhao, Herbert Jaeger, The error controlling algorithm for learning OOMs, Technical report 6, International University Bremen, 2007.