Low-Rank Spectral Learning

Alex Kulesza
Computer Science & Engineering
University of Michigan

N. Raj Rao
Electrical & Computer Engineering
University of Michigan

Satinder Singh
Computer Science & Engineering
University of Michigan

Abstract

Spectral learning methods have recently been proposed as alternatives to slow, non-convex optimization algorithms like EM for a variety of probabilistic models in which hidden information must be inferred by the learner. These methods are typically controlled by a rank hyperparameter that sets the complexity of the model; when the model rank matches the true rank of the process generating the data, the resulting predictions are provably consistent and admit finite sample convergence bounds. However, in practice we usually do not know the true rank, and, in any event, from a computational and statistical standpoint it is likely to be prohibitively large. It is therefore of great practical interest to understand the behavior of *low-rank* spectral learning, where the model rank is less than the true rank. Counterintuitively, we show that even when the singular values omitted by lowering the rank are arbitrarily small, the resulting prediction errors can in fact be arbitrarily large. We identify two distinct possible causes for this bad behavior, and illustrate them with simple examples. We then show that these two causes are essentially complete: assuming that they do not occur, we can prove that the prediction error is bounded in terms of the magnitudes of the omitted singular values. We argue that the assumptions necessary for this result are relatively realistic, making low-rank spectral learning a viable option for many applications.

1 INTRODUCTION

Traditionally, probabilistic models with hidden information (such as latent variable models) have been trained using iterative algorithms like expectation-maximization (EM) that attempt to traverse the non-convex likelihood landscape via local search [Dempster et al., 1977, McLachlan and Krishnan, 2007]. Though widely used, these algorithms are usually expensive to run, require many iterations, and typically cannot guarantee anything more than local improvement at each step [Wu, 1983].

These limitations might be seen as natural consequences of dealing with computationally hard learning problems; for instance, Terwijn [2002] showed that learning hidden Markov models (HMMs) is NP-hard. However, there has recently been significant interest in alternative learning methods [Cybenko and Crespi, 2011, Kontorovich et al., 2013]; in particular, the development of spectral learning techniques has suggested that relatively mild assumptions may in fact be sufficient to enable a class of computationally efficient and statistically consistent learning algorithms. These methods give closed-form parameter estimates, and so are usually simple to implement and fast to run. Spectral algorithms have been proposed for learning HMMs [Hsu et al., 2012] and other latent variable graphical models [Cohen et al., 2012, Anandkumar et al., 2012a,b, Parikh et al., 2011], as well as for predictive state representations [Boots et al., 2010b, Boots and Gordon, 2011], automata [Luque et al., 2012, Balle et al., 2011, 2013], and a number of other formalisms.

Fundamentally, spectral methods forego the likelihood optimization of EM and related methods[1] in favor of directly identifying hidden information in observed quantities. This process generally involves a singular value decomposition (SVD) on a moment matrix estimated from observed data; viewed as a canonical correlation analysis (CCA), the goal of this decomposition is to identify correlations between observations

Appearing in Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

[1] Balle et al. [2012] proposed some alternative optimization objectives inspired by spectral learning.

that must be explained by hidden information. For instance, the graphical structure of an HMM tells us that past observations are conditionally independent of future observations given the current state; thus, any correlations we observe between past and future observations must somehow reflect the hidden state. By identifying these correlations through SVD and then applying some clever algebraic manipulations, we can recover the original model when certain conditions are met.

On the other hand, one advantage of optimization-based approaches like EM is their inherent reasonableness; regardless of whether the assumptions of the model hold exactly in the real world, we can still hope to optimize for a good fit. An important practical question, then, is whether similar reasonable results hold for spectral learning.

We are particularly interested here in the question of *rank*: spectral methods depend on a rank hyperparameter that determines how many of the correlations estimated via SVD should be retained. As far as we are aware, all existing analyses of spectral learning assume that this model rank is chosen to exactly equal the true rank of the process generating the data; in other words, in the limit of infinite training data, all of the correlations corresponding to non-zero singular values will be kept. This assumption allows for proofs of statistical consistency and finite-sample bounds [Hsu et al., 2012, Boots et al., 2010a, Foster et al., 2012].

However, in practice we rarely know the correct rank; moreover, for any real-world process the rank is likely to be unbounded, or at least too large to be both computationally feasible (in terms of performing the SVD and other calculations) and statistically feasible (in terms of acquiring enough training data to reliably estimate the full set of correlations). Thus in reality we must usually resort to what we will call *low-rank* spectral learning, where the model rank is less than the true rank, and we merely take the correlations with the largest associated singular values. Indeed, this method has been previously suggested as a means of regularizing the complexity of spectral models [Boots et al., 2010b].

The low-rank approach makes intuitive sense, since we are used to treating the magnitudes of singular values as measures of "importance" for their associated singular vectors. However, we will show that this reasonable property does not generally hold for spectral learning. In particular, it is possible for the omission of arbitrarily small singular values to lead to arbitrarily large prediction errors. Moreover, in contrast to the statistical consistency of standard full-rank spectral learning, low-rank spectral learning can produce poor results even with an infinitely large training set, and even in cases where accurate low-rank models exist.

Concretely, we focus here on the simple and foundational HMM setup described by Hsu et al. [2012]. In this setting we identify two distinct possible sources of problematic behavior. The first is an incompatibility between the initial distribution over hidden states (from which training samples are drawn) and the long-term state dynamics; intuitively, a bad initial distribution can bias the learner in the wrong way. The second is a mapping from states to observations in which states do not look sufficiently distinct; this can dilute important information that would be apparent if the states were observed directly. We illustrate these scenarios with simple examples, each of which leads to large prediction errors despite omitting only small singular values.

We then complete the picture by showing that, in fact, these two cases fully characterize the settings in which low-rank spectral learning can fail. We prove that the prediction error of the low-rank model is bounded in terms of the largest omitted singular value whenever the initial state distribution is the stationary distribution and the observation model is well-conditioned. We argue that these assumptions are relatively realistic, in the sense that they can be observed or influenced by the practitioner in many cases.

2 BACKGROUND

In our setting, the world generates sequences of discrete observations $x_1, x_2, x_3, \ldots$ from the set $\{1, 2, \ldots, n\}$. The process generating these sequences is assumed to be a hidden Markov model (HMM) with states $\{1, 2, \ldots, m\}$. (Note that if m is allowed to be arbitrarily large then this is not actually a restriction at all.) We denote by y_t the hidden state at time t.

The parameters of the HMM include an initial state distribution $\pi \in \mathbb{R}^m$, $\Pr(y_1 = i) = \pi_i$, a transition matrix $T \in \mathbb{R}^{m \times m}$, $\Pr(y_{t+1} = i | y_t = j) = T_{ij}$, and an observation matrix $O \in \mathbb{R}^{n \times m}$, $\Pr(x_t = i | y_t = j) = O_{ij}$. We will make extensive use of the observable operators $A_x = TD_x$, where $D_x = \operatorname{diag}(O_{x\cdot})$ is the diagonal matrix whose ith diagonal entry is O_{xi}. Using these operators we can compute the joint probability of an observation sequence as

$$\Pr(x_1, x_2, \ldots, x_t) = \mathbf{1}^\top A_{x_{t:1}} \pi \ , \tag{1}$$

where $A_{x_{t:1}}$ is a shorthand denoting the product $A_{x_t} A_{x_{t-1}} \cdots A_{x_2} A_{x_1}$.

The goal of learning (for our purposes) will be to predict, from a training set of sampled observation sequences, the joint probabilities of all sequences of t observations for some finite constant t. In particular, we will consider the L_1 variational distance

$$\sum_{x_1,\ldots,x_t} \left| \widehat{\Pr}(x_1,\ldots,x_t) - \Pr(x_1,\ldots,x_t) \right| \ , \qquad (2)$$

where $\widehat{\Pr}$ denotes the predicted probability.

One possible approach to learning is to try and discover the original parameters π, T, and O; the standard expectation-maximization (EM) algorithm attempts this non-convex problem via alternating local optimization [Dempster et al., 1977]. However, EM can be slow in practice, and provides no guarantees regarding the quality of the final solution. Another approach would be to simply tally the observed counts for all sequences of length t, but this would require time, space, and training data exponential in t.

Instead, Hsu et al. [2012] showed that, assuming π is strictly positive and T and O are of rank m (full rank), a transformed parameterization of the HMM—sufficient to predict the desired joint probabilities, and more—is recoverable from quantities that depend only on observations:

$$\begin{aligned} P_1 &\in \mathbb{R}^n & [P_1]_i &= \Pr(x_1 = i) \\ P_{21} &\in \mathbb{R}^{n\times n} & [P_{21}]_{ij} &= \Pr(x_2 = i, x_1 = j) \qquad (3) \\ P_{3x1} &\in \mathbb{R}^{n\times n} & [P_{3x1}]_{ij} &= \Pr(x_3 = i, x_2 = x, x_1 = j) \ , \end{aligned}$$

where a P_{3x1} matrix is computed for each observation symbol x. Note that these statistics depend only on the distribution of the first three observations. They can also be written in terms of the HMM parameters:

$$\begin{aligned} P_1 &= O\pi \\ P_{21} &= OTD_\pi O^\top \qquad (4) \\ P_{3x1} &= OA_xTD_\pi O^\top \ , \end{aligned}$$

where $D_\pi = \mathrm{diag}(\pi)$.

The spectral model parameters are given in closed form in terms of the above P-statistics, as well as a matrix $U \in \mathbb{R}^{n\times m}$ with the property that $(U^\top O)$ is invertible. Typically, U is chosen to contain the m principal singular vectors of P_{21}, which always results in the desired property. The parameters are defined as

$$\begin{aligned} \boldsymbol{b}_1 &= U^\top P_1 \\ \boldsymbol{b}_\infty^\top &= P_1^\top (U^\top P_{21})^+ \qquad (5) \\ B_x &= U^\top P_{3x1}(U^\top P_{21})^+ \ , \end{aligned}$$

where again we have one matrix B_x for each possible observation x, and A^+ denotes the pseudoinverse of the matrix A. Equation (5) can be seen as the solution to a set of linear regression equations; for instance, B_x linearly transforms $U^\top P_{21}$ to $U^\top P_{3x1}$. Joint probability predictions are then computed from these parameters:

$$\widehat{\Pr}(x_1, x_2, \ldots, x_t) = \boldsymbol{b}_\infty^\top B_{x_{t:1}} \boldsymbol{b}_1 \ . \qquad (6)$$

Equation (6) parallels Equation (1); we can think of $\boldsymbol{b}_1$ as a transformed initial state vector, B_x as an observable update operator, and $\boldsymbol{b}_\infty^\top$ as a normalizer. Moreover, Hsu et al. [2012] showed that, when the P-statistics are exact, $\widehat{\Pr} = \Pr$.

In practice, when the exact P-statistics are not available, they are estimated by simply counting observations in the training set, and the model parameters are computed from these estimates using Equation (5). Hsu et al. [2012] showed that the resulting joint probability predictions are consistent in the limit of infinite data, and moreover that the size of the training set required to achieve a fixed level of accuracy is only polynomial in t.

3 CHALLENGES FOR LOW-RANK SPECTRAL LEARNING

Note that in order to compute the spectral parameters in Equation (5) we must first obtain U, which in turn requires knowing the number of HMM states m. As we have argued, in practice this number is likely to be unknown, and in any case impractically large. Instead, we will concern ourselves with *low-rank* spectral learning for HMMs, where the spectral projection $U \in \mathbb{R}^{n\times k}$ contains the k principal left singular vectors of P_{21} for some $k < m$. We refer to the singular values of P_{21} whose corresponding singular vectors are *not* included in U—that is, singular values that are not among the k largest—as the omitted singular values. The model parameters are computed from the low-rank U using Equation (5) as usual.

In order to simplify our discussion going forward, we will ignore errors that arise from having a limited training set and instead assume that we have perfect estimates of the P-statistics. This means that we can use the closed-form expressions in Equation (4). Though unrealistic, assuming an unbounded training set allows us to isolate the effects of learning a low-rank model from finite-sample convergence issues. (Whether finite-sample issues would compound the challenges of low-rank spectral learning in an interesting way—or, perhaps, alleviate them—remains an interesting and open question.)

Of course, in practice, having a finite training set is an important motivation for the use of low-rank learning: accurately estimating the singular vectors of P_{21} with small corresponding singular values is problematic precisely because very large quantities of data are needed to do so [Benaych-Georges and Nadakuditi, 2012]. Indeed, existing finite sample bounds typically have a term that grows like $O(1/\sigma_{\min}^4)$, where $\sigma_{\min}$ is the smallest nonzero singular value of P_{21} (or equiva-

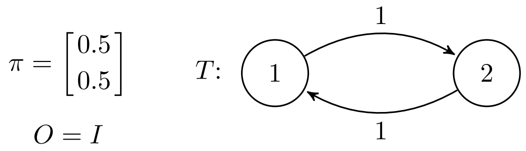

Figure 1: A simple two-state HMM.

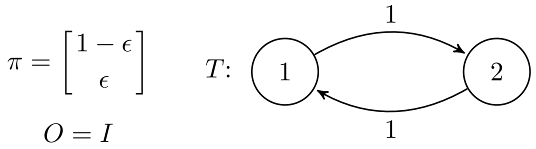
$\pi = \begin{bmatrix} 1-\epsilon \\ \epsilon \end{bmatrix}$
T:
1
1
1
2
$O = I$

Figure 2: A modified two-state HMM.

lent) [Hsu et al., 2012, Boots et al., 2010a, Foster et al., 2012].

3.1 Restricting π

We might assume, intuitively, that a choice of k where the omitted singular values are small enough cannot cause large errors. This turns out not to be true. We present a series of three examples illustrating the depth of the problem.

First example. Consider the simple two-state, two-observation HMM in Figure 1. (T is depicted as an automaton, with states drawn as circles and probabilities on the transition arrows.) O is the identity matrix, so in effect this is a (non-hidden) Markov model. It is clear by inspection that this HMM produces only the alternating observation sequences $1, 2, 1, 2, \ldots$ and $2, 1, 2, 1, \ldots$, and each occurs with probability 50%. We can easily compute

$$P_{21} = \left[\begin{array}{cc} 0 & 0.5 \\ 0.5 & 0 \end{array} \right] , \tag{7}$$

which has singular values $(0.5, 0.5)$. The top singular vector can be any unit vector, but if a small amount of noise is added it will be an elementary basis vector such as $[0\ 1]^\top$. It is easy to compute that rank-one spectral learning in this case yields $B_1 = B_2 = 0$; therefore the model predicts zero probability for every sequence, and the L_1 variational distance is 1 for all t.

This is not terribly surprising: the large singular values of P_{21} are a clear suggestion that lowering the rank might result in a poor approximation. But does this implication hold in reverse? That is, do small singular values imply that a low-rank model will be a good fit? The next example shows that the answer, in general, is negative.

Second example. Figure 2 depicts a slight modification to the HMM in Figure 1. The only change is to π; here ϵ is some small positive number. Whereas before the two feasible sequences had equal probability, the sequence $1, 2, 1, 2, \ldots$ is now observed almost all of the time. We have

$$P_{21} = \left[\begin{array}{cc} 0 & \epsilon \\ 1-\epsilon & 0 \end{array} \right] , \tag{8}$$

with singular values $(1-\epsilon, \epsilon)$. This time we might reasonably suppose that the second singular vector is unimportant, given its small associated singular value. And yet, again, simple computations show that a rank-one spectral model yields $B_1 = B_2 = 0$ and gives trivial predictions. This means that there can in general be no "safe" threshold for pruning the singular values of P_{21}; an arbitrarily small singular value might still be crucially important. In this case, a rank-two model recovers the process perfectly, while a rank-one model is totally uninformative.

One way to view this result is that, though we have technically met the spectral learning condition that $\pi > 0$, we have "barely" met it by setting $\pi_2 = \epsilon$. In the same way that spectral learning fails when an element of π is equal to 0, we should somehow expect increasing difficulty as an element of π *approaches* zero. This is true, in the sense that if the conditions from Section 2 are met with margin then the singular values of P_{21} cannot get too small. (It is not, however, sufficient to simply ensure that the entries of π are not too small.) Nonetheless, this is not a satisfying resolution, since assuming that the singular values of P_{21} do not get too small is equivalent to assuming that its rank, m, is easily obtained from a finite sample, which we have argued is not feasible in the first place.

We can, however, provide at least one sound reason for the poor performance of the rank-one spectral model here: we have designed an HMM that cannot be effectively approximated by *any* rank-one model, since the alternating pattern of observations is fundamentally stateful. Perhaps, even if singular values fail to convey the value of increasing model rank, low-rank spectral learning will still recover a model that is near-optimal (with respect to some reasonable objective) within the class of low-rank models. Unfortunately, the next example shows that this cannot be true either.

Third example. Compared to Figure 2, the HMM in Figure 3 adds a "dummy" state that always transitions to itself and allocates to it a small positive amount δ of the initial probability mass. This HMM generally behaves the same as its predecessor, but with probability δ it produces only 3s. We now have

$$P_{21} = \left[\begin{array}{ccc} 0 & \epsilon & 0 \\ 1-\epsilon-\delta & 0 & 0 \\ 0 & 0 & \delta \end{array} \right] , \tag{9}$$

$$\pi = \begin{bmatrix} 1-\epsilon-\delta \\ \epsilon \\ \delta \end{bmatrix}$$

$$O = I$$

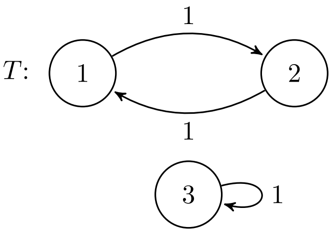

Figure 3: A problematic three-state HMM.

$$\pi = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}^\top$$

$$O = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1-\epsilon \\ 0 & 0 & \epsilon \end{bmatrix}$$

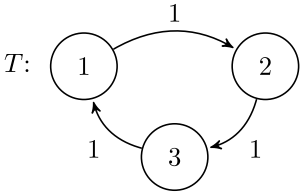

Figure 4: An HMM with $T\pi = \pi$.

with singular values $(1-\epsilon-\delta, \epsilon, \delta)$. By construction, there exists a rank-two model that gives arbitrarily good predictions as $\delta \to 0$, obtained by simply ignoring the third state. Indeed, when $\epsilon > \delta$ rank-two spectral learning recovers this result, giving nearly perfect predictions for all lengths t. When $\delta > \epsilon$, however, the alternating pattern is "masked" by the dummy state. The learner chooses to allocate its representational power poorly, and the result is an L_1 variational distance approaching 1.

The last two examples have something in common: they both rely on an initial distribution π that is biased toward the first state, while in the long run states 1 and 2 are equally likely. Another way to say this is that the first few time steps, from which the learner computes all of its parameters, do not accurately reflect the long-term dynamics of the process. We might hope to ameliorate this problem by averaging the P-statistics across all time steps in a very long observation sequence; equivalently, we can require that the initial distribution π is also the stationary distribution, $T\pi = \pi$. Indeed, the examples shown so far are well-behaved under this modification.

3.2 Restricting O

Does fixing π to the stationary distribution solve all of our problems? Unfortunately, no. Consider the HMM in Figure 4. This process generates a repeating sequence $1, 2, 2, \ldots$ or $2, 2, 1, \ldots$ or $2, 1, 2, \ldots$ with equal probability; very rarely, a second consecutive 2 will be replaced with a 3. In this case,

$$P_{21} = \begin{bmatrix} 0 & \frac{1-\epsilon}{3} & \frac{\epsilon}{3} \\ \frac{1}{3} & \frac{1-\epsilon}{3} & 0 \\ 0 & \frac{\epsilon}{3} & 0 \end{bmatrix} , \tag{10}$$

and from the third column we can see that as $\epsilon \to 0$ the smallest singular value of P_{21} will approach zero as well. At the same time, it is not possible to implement repeating patterns of length three with only two states, and the rank-two spectral model always gives large errors for $t \geq 4$. While this is another case where no good low-rank model exists, it is straightforward to add a dummy state as in the third example; then a good rank-three model exists, but will not be found.

As this example demonstrates, merely assuming $T\pi = \pi$ is not sufficient to guarantee successful low-rank learning; an observation matrix O with small singular values can make dissimilar states appear similar. This is a second possible source of poor performance.

As we will prove in the next section, it turns out that no special requirements on T are needed for low-rank spectral learning (other than its compatibility with π). Intuitively, this is because T controls the long-term dynamics of the HMM; if T has a small singular value, it is because there is at least one state that can be well-approximated by a combination of other states. In contrast, small singular values in O may cause states to appear redundant at the current time step, when in fact they have significantly different long-term dynamics.

4 ERROR BOUND

We have seen that low-rank spectral learning can fail when the initial distribution π is not matched to the transition matrix T, or when the observation matrix O is poorly conditioned. In this section we show that, essentially, these are the *only* ways in which low-rank spectral learning can fail. We will prove the following theorem.

Theorem 1. *Let O, T, and π define a hidden Markov model with $m \geq 4$ states and n observations satisfying* $\mathrm{rank}(O) = m$*, $\pi > 0$, and $T\pi = \pi$. (That is, the initial distribution is also the stationary distribution.) Let P_1, P_{21}, P_{3x1}, $\boldsymbol{b}_1$, $\boldsymbol{b}_\infty^\top$, and B_x be defined as in Sections 2 and 3, and let U be the $n \times k$ matrix containing the top k left singular vectors of P_{21}. Let σ_{k+1} denote the $(k+1)$th largest singular value of P_{21}; i.e., σ_{k+1} is the largest omitted singular value. We have the following bound for any t:*

$$\sum_{x_1,\ldots,x_t} \left| \widehat{\Pr}(x_1,\ldots,x_t) - \Pr(x_1,\ldots,x_t) \right| \leq \sqrt{m} \left(\sigma_{\min}^{-1}(O)\sqrt{m}\right)^{t+3} \sigma_{k+1} , \tag{11}$$

where $\sigma_{\min}(O)$ is the smallest singular value of O.

Theorem 1 says that, when $T\pi = \pi$ and $\sigma_{\min}(O)$ is not too small, the L_1 variational distance between the true and estimated distributions for a given t is controlled primarily by σ_{k+1}. The result is independent of n, making it attractive for applications where the set of possible observations might be very large.

In prior work, Siddiqi et al. [2010] observed that when T is low-rank, the resulting zeros in the spectrum of P_{21} can be omitted without affecting the recovery of the true model. Theorem 1 takes this idea further, showing that under the right conditions even non-zero singular values can be omitted with only a bounded effect on the resulting predictions.

Note that the bound in Theorem 1 is exponential in t, in contrast the the bounds of Hsu et al. [2012], which depend on t only polynomially. However, these bounds characterize fundamentally different sources of error: Hsu et al. [2012] consider the effects of finite-sample estimation, whereas we are interested in low-rank approximation when the P-statistics are exact. Moreover, when $\sigma_{\min}(O) > \sqrt{m}/n$, the exponential term in Theorem 1 at least grows more slowly than the number of possible observation sequences, so the expected error of a single t-length prediction goes to zero.

We state and prove three important lemmas before proceeding to the proof of Theorem 1. Our first lemma characterizes the model's predictions in terms of a low-rank operator that is repeatedly applied at each time step. This perspective is intuitive, and motivates the rest of our analysis.

Lemma 1. *Define the low-rank operator $M = O^+UU^\top O$, and let $\hat{A}_x = MA_x$ and $\hat{\pi} = M\pi$. For the low-rank spectral learning parameters $\boldsymbol{b}_1$, $\boldsymbol{b}_\infty^\top$, and B_x defined in Section 2, we have, for all t and all observation sequences $x_1, \ldots, x_t$:*

$$\boldsymbol{b}_\infty^\top B_{x_{t:1}} \boldsymbol{b}_1 = \mathbf{1}^\top \hat{A}_{x_{t:1}} \hat{\pi} \ .$$

Proof. We first show by induction that

$$B_{x_{t:1}} \boldsymbol{b}_1 = U^\top O A_{x_t} \hat{A}_{x_{t-1:1}} \hat{\pi} \ . \tag{12}$$

When $t = 1$, we have

$$\begin{aligned}
B_{x_1} \boldsymbol{b}_1 &= U^\top P_{3x_1 1} (U^\top P_{21})^+ U^\top P_1 && (13)\\
&= U^\top O A_{x_1} T D_\pi O^\top (U^\top P_{21})^+ U^\top O \pi && (14)\\
&= U^\top O A_{x_1} O^+ O T D_\pi O^\top (U^\top P_{21})^+ U^\top O \pi && (15)\\
&= U^\top O A_{x_1} O^+ P_{21} (U^\top P_{21})^+ U^\top O \pi \ , && (16)
\end{aligned}$$

where we use the fact that $O^+O = I$ (since $\text{rank}(O) = m$, by assumption). Since U consists of left singular vectors of P_{21}, $P_{21}(U^\top P_{21})^+ = U$. So the above is equal to

$$U^\top O A_{x_1} O^+ UU^\top O \pi = U^\top O A_{x_1} \hat{\pi} \ . \tag{17}$$

The inductive step follows from the same argument.

Finally, we have

$$\begin{aligned}
\boldsymbol{b}_\infty^\top U^\top O A_{x_t} &= P_1^\top (U^\top P_{21})^+ U^\top O A_{x_t} && (18)\\
&= \pi^\top O^\top (U^\top P_{21})^+ U^\top O A_{x_t} && (19)\\
&= \mathbf{1}^\top D_\pi O^\top (U^\top P_{21})^+ U^\top O A_{x_t} && (20)\\
&= \mathbf{1}^\top T D_\pi O^\top (U^\top P_{21})^+ U^\top O A_{x_t} && (21)\\
&= \mathbf{1}^\top O^+ P_{21} (U^\top P_{21})^+ U^\top O A_{x_t} && (22)\\
&= \mathbf{1}^\top O^+ UU^\top O A_{x_t} && (23)\\
&= \mathbf{1}^\top \hat{A}_{x_t} \ , && (24)
\end{aligned}$$

since T is left stochastic. □

Our next lemma concerns the approximation error introduced by the low-rank operator M if it is applied only at the very last time step. Intuitively, this error might be large because U is derived from P_{21} and therefore biased by the initial distribution π. However, when $T\pi = \pi$, the long-term state distribution mirrors the initial distribution, and we can show that the error does not get too large.

Lemma 2. *Under the conditions of Theorem 1, for any t,*

$$\sum_{x_1, \ldots, x_t} \left\| \hat{A}_{x_t} A_{x_{t-1:1}} \pi - A_{x_{t:1}} \pi \right\|_2 \leq \sigma_{\min}^{-2}(O) m \sigma_{k+1} \ .$$

Proof. We begin by expanding $\hat{A}_{x_t}$ and A_{x_t}, and then introduce π as a "reference" distribution:

$$\begin{aligned}
&\sum_{x_1, \ldots, x_t} \left\| \hat{A}_{x_t} A_{x_{t-1:1}} \pi - A_{x_{t:1}} \pi \right\|_2 && (25)\\
&= \sum_{x_1, \ldots, x_t} \left\| M T D_{x_t} A_{x_{t-1:1}} \pi - T D_{x_t} A_{x_{t-1:1}} \pi \right\|_2 && (26)\\
&= \sum_{x_1, \ldots, x_t} \left\| \left(M T D_\pi - T D_\pi \right) D_\pi^{-1} D_{x_t} A_{x_{t-1:1}} \pi \right\|_2 \ .
\end{aligned}$$

Note that the inverse D_π^{-1} exists since $\pi > 0$. We next use the fact that $\left\| A\boldsymbol{x} \right\|_2 \leq \left\| A \right\|_2 \left\| \boldsymbol{x} \right\|_2$ for any matrix A and vector $\boldsymbol{x}$. (Here $\left\| A \right\|_2$ denotes the matrix norm induced by the L_2 vector norm, which is equal to the maximum singular value of A.) This bounds the above by

$$\begin{aligned}
&\left\| M T D_\pi - T D_\pi \right\|_2 \sum_{x_1, \ldots, x_t} \left\| D_\pi^{-1} D_{x_t} A_{x_{t-1:1}} \pi \right\|_2 && (27)\\
&\leq \left\| M T D_\pi - T D_\pi \right\|_2 \sum_{x_1, \ldots, x_t} \left\| D_\pi^{-1} D_{x_t} A_{x_{t-1:1}} \pi \right\|_1 \ ,
\end{aligned}$$

since $\left\| \boldsymbol{x} \right\|_2 \leq \left\| \boldsymbol{x} \right\|_1$ for any vector $\boldsymbol{x}$.

We first address the sum. Note that $D_\pi^{-1} D_{x_t} A_{x_{t-1:1}} \pi$ has only non-negative entries, so we can replace the norm by a simple sum:

$$\sum_{x_1, \ldots, x_t} \left\| D_\pi^{-1} D_{x_t} A_{x_{t-1:1}} \pi \right\|_1 \tag{28}$$

$$
\begin{aligned}
&= \sum_{x_1,\ldots,x_t} \mathbf{1}^\top D_\pi^{-1} D_{x_t} A_{x_{t-1:1}} \pi && (29)\\
&= \sum_{i=1}^{m} \frac{1}{\pi_i} \sum_{x_1,\ldots,x_t} \left[D_{x_t} A_{x_{t-1:1}} \pi\right]_i \ . && (30)
\end{aligned}
$$

Recall that $\left[A_{x_{t-1:1}}\pi\right]_i$ is the (true) joint probability of observing $x_1, \ldots, x_{t-1}$ and ending up in state i at time t, thus $\left[D_{x_t} A_{x_{t-1:1}}\pi\right]_i$ is the joint probability of observing $x_1, \ldots, x_{t-1}$, ending up in state i at time t, and then observing x_t. Summed over all sequences $x_1, \ldots, x_t$, this is just the probability of being in state i at time t, which, by definition, is the stationary probability π_i. Thus the above is equal to

$$
\sum_{i=1}^{m} \frac{1}{\pi_i} \pi_i = m \ . \tag{31}
$$

Now consider the term outside the sum; we have

$$
\begin{aligned}
&\|MTD_\pi - TD_\pi\|_2 && (32)\\
&= \left\|O^+ UU^\top OTD_\pi - TD_\pi\right\|_2 && (33)\\
&\leq \left\|O^+\right\|_2 \left\|UU^\top OTD_\pi O^\top - OTD_\pi O^\top\right\|_2 \left\|(O^\top)^+\right\|_2 && (34)\\
&= \sigma_{\min}^{-2}(O) \left\|UU^\top P_{21} - P_{21}\right\|_2 \ , && (35)
\end{aligned}
$$

since $\|O^+\|_2 = \sigma_{\min}^{-1}(O)$. Because $UU^\top$ is a projection onto the top k singular vectors of P_{21}, $\left\|UU^\top P_{21} - P_{21}\right\|_2$ is exactly σ_{k+1}. Combining, we have the desired result. □

The final lemma uses Lemma 2 and an inductive argument to establish a bound on the summed L_2 distances between the true and approximated belief states after t time steps. With this in hand, the proof of Theorem 1 will be straightforward.

Lemma 3. *Under the conditions of Theorem 1, for any t,*

$$
\sum_{x_1,\ldots,x_t} \left\|\hat{A}_{x_{t:1}} \hat{\pi} - A_{x_{t:1}} \pi\right\|_2 \leq \sigma_{k+1} \sum_{r=2}^{t+2} \left(\sigma_{\min}^{-1}(O)\sqrt{m}\right)^r \ .
$$

Proof. We will prove the claim by induction. As a base case, let $t = 0$. Note that

$$
\begin{aligned}
\pi &= T\pi && (36)\\
&= TD_\pi \mathbf{1} && (37)\\
&= TD_\pi O^\top \mathbf{1} \ , && (38)
\end{aligned}
$$

since $T\pi = \pi$ by assumption and $O^\top$ is right stochastic; thus

$$
\|\hat{\pi} - \pi\|_2 = \left\|O^+ UU^\top OTD_\pi O^\top \mathbf{1} - TD_\pi O^\top \mathbf{1}\right\|_2 \ . \tag{39}
$$

Proceeding similarly to Lemma 2, this is bounded by

$$
\begin{aligned}
&\left\|O^+\right\|_2 \left\|UU^\top OTD_\pi O^\top \mathbf{1} - OTD_\pi O^\top \mathbf{1}\right\|_2 && (40)\\
&\quad\leq \sigma_{\min}^{-1}(O)\sqrt{m} \left\|UU^\top P_{21} - P_{21}\right\|_2 && (41)\\
&\quad= \sigma_{\min}^{-1}(O)\sqrt{m}\,\sigma_{k+1} \ , && (42)
\end{aligned}
$$

since $\|\mathbf{1}\|_2 = \sqrt{m}$ when $\mathbf{1}$ is m-dimensional. Since O is stochastic, $\sigma_{\min}(O) \leq 1$ and the above is bounded by

$$
\sigma_{\min}^{-2}(O) m \sigma_{k+1} \ . \tag{43}
$$

Thus the claim holds for $t = 0$.

Now suppose, inductively, that the claim holds for $t-1$. To prove the claim for t, we first apply the triangle inequality to split the sum into more manageable pieces:

$$
\begin{aligned}
&\sum_{x_1,\ldots,x_t} \left\|\hat{A}_{x_{t:1}} \hat{\pi} - A_{x_{t:1}} \pi\right\|_2 && (44)\\
&\quad\leq \sum_{x_1,\ldots,x_t} \left\|\hat{A}_{x_{t:1}} \hat{\pi} - \hat{A}_{x_t} A_{x_{t-1:1}} \pi\right\|_2 \\
&\quad\quad+ \sum_{x_1,\ldots,x_t} \left\|\hat{A}_{x_t} A_{x_{t-1:1}} \pi - A_{x_{t:1}} \pi\right\|_2 \ . && (45)
\end{aligned}
$$

Consider the first sum:

$$
\begin{aligned}
&\sum_{x_1,\ldots,x_t} \left\|\hat{A}_{x_t} \hat{A}_{x_{t-1:1}} \hat{\pi} - \hat{A}_{x_t} A_{x_{t-1:1}} \pi\right\|_2 && (46)\\
&\leq \sum_{x_1,\ldots,x_t} \left\|O^+ UU^\top\right\|_2 \left\|OA_{x_t}\left(\hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi\right)\right\|_2 \\
&\leq \sigma_{\min}^{-1}(O) \sum_{x_1,\ldots,x_t} \left\|OA_{x_t}\left(\hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi\right)\right\|_1 \ ,
\end{aligned}
$$

since $\left\|UU^\top\right\|_2 = 1$. Note that in the last step we replaced the L_2 norm with the larger L_1 norm. Now, since OA_{x_t} has non-negative entries,

$$
\begin{aligned}
&\left\|OA_{x_t}\left(\hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi\right)\right\|_1 && (47)\\
&\quad= \sum_{i=1}^{m} \left|\sum_{j=1}^{m} [OA_{x_t}]_{ij} \left[\hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi\right]_j\right| && (48)\\
&\quad\leq \sum_{i=1}^{m} \sum_{j=1}^{m} [OA_{x_t}]_{ij} \left|\left[\hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi\right]_j\right| && (49)\\
&\quad= \mathbf{1}^\top OA_{x_t} \left|\hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi\right| && (50)\\
&\quad= \mathbf{1}^\top A_{x_t} \left|\hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi\right| \ , && (51)
\end{aligned}
$$

where the absolute value of a vector is interpreted element-wise. Thus Equation (46) is bounded above by

$$
\sigma_{\min}^{-1}(O) \sum_{x_1,\ldots,x_t} \mathbf{1}^\top A_{x_t} \left|\hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi\right| \tag{52}
$$

$$= \sigma_{\min}^{-1}(O) \sum_{x_1,\ldots,x_{t-1}} \mathbf{1}^\top T \left| \hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi \right| \tag{53}$$

$$= \sigma_{\min}^{-1}(O) \sum_{x_1,\ldots,x_{t-1}} \left\| \hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi \right\|_1 \tag{54}$$

$$\leq \sigma_{\min}^{-1}(O) \sqrt{m} \sum_{x_1,\ldots,x_{t-1}} \left\| \hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi \right\|_2 ,$$

since $\sum_x A_x = T$, T is left stochastic, and $\|\boldsymbol{x}\|_1 \leq \sqrt{m}\,\|\boldsymbol{x}\|_2$ for any vector $\boldsymbol{x} \in \mathbb{R}^m$.

The second sum in Equation (45) is directly bounded by Lemma 2. Combining and using the shorthand $c = \sigma_{\min}^{-1}(O)\sqrt{m}$, we have

$$\sum_{x_1,\ldots,x_t} \left\| \hat{A}_{x_{t:1}} \hat{\pi} - A_{x_{t:1}} \pi \right\|_2 \tag{55}$$

$$\leq c \sum_{x_1,\ldots,x_{t-1}} \left\| \hat{A}_{x_{t-1:1}} \hat{\pi} - A_{x_{t-1:1}} \pi \right\|_2 + c^2 \sigma_{k+1} \tag{56}$$

$$\leq c\sigma_{k+1} \sum_{r=2}^{t+1} c^r + c^2 \sigma_{k+1} \tag{57}$$

$$= \sigma_{k+1} \sum_{r=3}^{t+2} c^r + c^2 \sigma_{k+1} \tag{58}$$

$$= \sigma_{k+1} \sum_{r=2}^{t+2} c^r , \tag{59}$$

where we apply the inductive hypothesis and simplify. □

We are now ready to prove the main theorem.

Proof of Theorem 1. Expanding the definition of $\Pr(\cdot)$ and applying Lemma 1, we have

$$\sum_{x_1,\ldots,x_t} \left| \widehat{\Pr}(x_1,\ldots,x_t) - \Pr(x_1,\ldots,x_t) \right|$$

$$= \sum_{x_1,\ldots,x_t} \left| \mathbf{1}^\top \hat{A}_{x_{t:1}} \hat{\pi} - \mathbf{1}^\top A_{x_{t:1}} \pi \right| . \tag{60}$$

Since $|\mathbf{1}^\top \boldsymbol{x}| \leq \|\boldsymbol{x}\|_1 \leq \sqrt{m}\,\|\boldsymbol{x}\|_2$ for $\boldsymbol{x} \in \mathbb{R}^m$, this is bounded above by

$$\sqrt{m} \sum_{x_1,\ldots,x_t} \left\| \hat{A}_{x_{t:1}} \hat{\pi} - A_{x_{t:1}} \pi \right\|_2 \tag{61}$$

$$\leq \sqrt{m}\sigma_{k+1} \sum_{r=2}^{t+2} \left(\sigma_{\min}^{-1}(O)\sqrt{m} \right)^r , \tag{62}$$

applying Lemma 3. All that remains is some simplification. Since $\sqrt{m} \geq 2$ by assumption and $\sigma_{\min}^{-1}(O) \geq 1$, we can bound the sum by $(\sigma_{\min}^{-1}(O)\sqrt{m})^{t+3}$. This gives the bound

$$\sqrt{m} \left(\sigma_{\min}^{-1}(O)\sqrt{m} \right)^{t+3} \sigma_{k+1} . \tag{63}$$

□

5 DISCUSSION

The spectral learning result of Hsu et al. [2012] relies on three basic assumptions: $\pi > 0$, $\operatorname{rank}(O) = m$, and $\operatorname{rank}(T) = m$. Theorem 1 adds to these the stationarity condition $T\pi = \pi$ (and drops the rank condition on T, as discussed in Section 4). Is the stationarity condition realistic? While in practice we usually have no control over the initial distribution from which our training data are generated, applications where training and test data take the form of long observation sequences generated from continuous runs of the process allow us to satisfy the condition by computing P-statistics from the entire data stream and not just the first three time steps, assuming that the mixing rate is sufficiently high.[2] For instance, suppose we wish to model weather patterns using a set of daily weather reports from the past fifty years. The first few days in the training set might have had unusual weather merely by chance, but for all practical purposes a day sampled at random reflects the stationary state of the world. On the other hand, in natural language applications training examples are often individual sentences whose first words may behave in ways that do not reflect the long-term dynamics of the underlying model. For such applications the limitations of low-rank spectral learning are likely to be of greater concern.

Although not stated as an explicit assumption, to give useful bounds Theorem 1 also requires a second property: O must be well-conditioned. For some applications this may be achievable, particularly if the practitioner can influence the observation space, for instance, by adding sensors; if each state has a unique observation profile then O will tend not to have small singular values. Alternatively, the conditioning of O can potentially be improved by treating several consecutive observations as a single multi-observation. In Figure 4, for example, the distribution over the next *two* observations distinguishes all three states. This technique is widely used in the literature on predictive state representations [Wolfe et al., 2005, Boots et al., 2010b] and spectral learning of automata [Balle et al., 2013]. It seems plausible that such a procedure might provably address the problem of a poorly conditioned O, but this remains an open question for future work.

Acknowledgements

This work was supported by NSF grant 1319365. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

[2] Note that, while a fast mixing rate may play a role in getting accurate estimates of the P-statistics, Theorem 1 shows it is not otherwise required for learning.

References

Anima Anandkumar, Dean Foster, Daniel Hsu, Sham Kakade, and Yi-Kai Liu. A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 25*, pages 926–934, 2012a.

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden Markov models. *arXiv preprint arXiv:1203.0683*, 2012b.

Borja Balle, Ariadna Quattoni, and Xavier Carreras. A spectral learning algorithm for finite state transducers. In *Machine Learning and Knowledge Discovery in Databases*, pages 156–171. Springer, 2011.

Borja Balle, Ariadna Quattoni, and Xavier Carreras. Local loss optimization in operator models: A new insight into spectral learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1879–1886, 2012.

Borja Balle, Xavier Carreras, Franco M Luque, and Ariadna Quattoni. Spectral learning of weighted automata. *Machine Learning*, pages 1–31, 2013.

Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

Byron Boots and Geoffrey J Gordon. An online spectral learning algorithm for partially observable nonlinear dynamical systems. In *Proceedings of the 25th National Conference on Artificial Intelligence (AAAI-2011)*, 2011.

Byron Boots, Sajid M Siddiqi, Geoffrey Gordon, and Alex Smola. Hilbert space embeddings of hidden Markov models. In *Proc. 27th Intl. Conf. on Machine Learning (ICML)*, 2010a.

Byron Boots, Sajid M Siddiqi, and Geoffrey J Gordon. Closing the learning-planning loop with predictive state representations. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 1369–1370. International Foundation for Autonomous Agents and Multiagent Systems, 2010b.

Shay B Cohen, Karl Stratos, Michael Collins, Dean P Foster, and Lyle Ungar. Spectral learning of latent-variable PCFGs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 223–231. Association for Computational Linguistics, 2012.

George Cybenko and Valentino Crespi. Learning hidden Markov models using nonnegative matrix factorization. *Information Theory, IEEE Transactions on*, 57 (6):3963–3970, 2011.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

Dean P Foster, Jordan Rodu, and Lyle H Ungar. Spectral dimensionality reduction for HMMs. *arXiv preprint arXiv:1203.6130*, 2012.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78 (5):1460–1480, 2012.

Aryeh Kontorovich, Boaz Nadler, and Roi Weiss. On learning parametric-output HMMs. In *Proceedings of The 30th International Conference on Machine Learning*, pages 702–710, 2013.

Franco M Luque, Ariadna Quattoni, Borja Balle, and Xavier Carreras. Spectral learning for non-deterministic dependency parsing. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 409–419. Association for Computational Linguistics, 2012.

Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 2007.

A Parikh, L Song, and Eric P Xing. A spectral algorithm for latent tree graphical models. In *Proceedings of The 28th International Conference on Machine Learning*, 2011.

Sajid M Siddiqi, Byron Boots, and Geoffrey J Gordon. Reduced-rank hidden Markov models. In *International Conference on Artificial Intelligence and Statistics*, pages 741–748, 2010.

Sebastiaan A Terwijn. On the learnability of hidden Markov models. In *Grammatical Inference: Algorithms and Applications*, pages 261–268. Springer, 2002.

Britton Wolfe, Michael R James, and Satinder Singh. Learning predictive state representations in dynamical systems without reset. In *Proceedings of the 22nd international conference on Machine learning*, pages 980–987. ACM, 2005.

CF Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.