An Exploration of Low-Rank Spectral Learning

Alex Kulesza KULESZA@UMICH.EDU
N. Raj Rao RAJNRAO@UMICH.EDU
Satinder Singh BAVEJA@UMICH.EDU
University of Michigan Ann Arbor, MI 48109 USA

Abstract

Spectral learning methods have recently been proposed for a variety of probabilistic models. These methods typically involve a rank hyperparameter that controls the complexity of the model; when it is set to match the true rank of the process generating the training data, the resulting estimate is provably consistent and admits finite sample convergence bounds. However, in practice we usually do not know the true rank, and in any case, from a computational 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. In this paper we show that several appealing and intuitive hypotheses turn out to be false, including that the error of the learned model is bounded in some way by the magnitudes of the omitted singular values, and that spectral methods are guaranteed to converge to good low-rank models if they exist. We present a series of simple examples illustrating these negative results, and discuss their implications. Finally, we present some synthetic results suggesting that there may be special cases where these problems can be controlled.

1. Introduction

Spectral learning methods have become popular alternatives to slow, non-convex algorithms like EM for models in which hidden information must be inferred by the learner. In spectral learning, such information is typically discovered through the singular value decomposition of a specially constructed correlation matrix; this process gives the method its name, and can be seen as a form of canonical correlation analysis (CCA). Because these calculations have a closed form, spectral methods are typically much faster than algorithms that, like EM, attempt to iteratively optimize an objective function. However, less is understood about their behavior in settings that do not meet the assumptions of existing analysis. In this paper we are particularly interested in the issue of learning low-rank spectral models, which we describe in more detail in Section 3. We begin with some background.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

2. Background

Spectral learning techniques have been developed for a variety of latent variable graphical models (Hsu et al., 2012; Cohen et al., 2012; Anandkumar et al., 2012; Parikh et al., 2011), predictive state representations (Boots et al., 2010b; Boots & Gordon, 2011), automata (Luque et al., 2012; Balle et al., 2011; Balle & Mohri, 2012), and many other settings. We focus here on the simple and widely known HMM setup described by Hsu et al. (2012); however, our aim is to address broad questions about spectral learning that arise for many different models.

The basic setup is as follows. The world produces sequences of discrete observations $x_1, x_2, x_3, \ldots$ from the set $\{1, 2, \ldots, n\}$. The process generating these sequences is a hidden Markov model (HMM) with states $\{1, 2, \ldots, m\}$, where the hidden state at time t is given by y_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 $O \in \mathbb{R}^{n \times m}$, $\Pr(x_t = i | y_t = j) = O_{ij}$. Defining the observable operators $A_x = T\mathrm{diag}(O_{x\cdot})$, we can write the joint probability of an observation sequence as

$$\Pr(x_1, x_2, \ldots, x_t) = \mathbf{1}^\top A_{x_t} \cdots A_{x_2} A_{x_1} \pi \; . \qquad (1)$$

The goal of learning (for our purposes) is 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

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

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

One possible approach is to try and discover the original parameters π, T, and O; the standard EM algorithm (Dempster et al., 1977) attempts this non-convex problem via alternating local optimization. However, EM can be slow in practice, and provides no guarantees regarding the quality of the final solution.

Instead, Hsu et al. (2012) showed that, under the conditions that π is strictly positive and T and O are of rank m, a transformed parameterization of the HMM—sufficient to predict the desired joint probabilities, and more—is recoverable from quantities that can be computed using only visible observations:

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

where a P_{3x1} matrix is computed for each observation symbol x. The spectral parameters are given in terms of 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}. The parameters are:

$$\begin{aligned} b_1 &= U^\top P_1 \\ b_\infty^\top &= P_1^\top (U^\top P_{21})^+ \\ B_x &= U^\top P_{3x1} (U^\top P_{21})^+ , \end{aligned} \qquad (3)$$

where again we have B_x for each observation x.

In practice, when exact P-statistics are not available, they are simply estimated by counting. Hsu et al. (2012) showed that the resulting joint probability estimates

$$\Pr(x_1, x_2, \ldots, x_t) \approx b_\infty^\top B_{x_t} \cdots B_{x_2} B_{x_1} b_1 \qquad (4)$$

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. Low-rank spectral learning

As far as we are aware, existing analyses of spectral learning assume that the number of states m (or an equivalent complexity measure, depending on the model) is known to the learner. This allows for proofs of statistical consistency and finite-sample bounds. However, in practice we rarely know the correct value of m; moreover, for any real process it is likely to be unbounded, or at least too large to be computationally feasible. Thus we must usually resort to what we will call *low-rank* spectral learning, where the spectral projection $U \in \mathbb{R}^{n\times k}$ contains the k principal left singular vectors of P_{21} (or equivalent) for some $k < m$. Indeed, this method has been previously suggested as a means of regularizing the complexity of spectral models (Boots et al., 2010b).

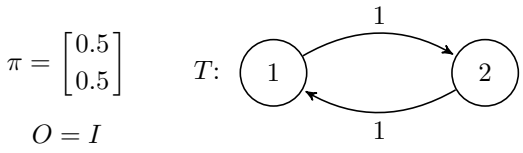

Figure 1. A simple two-state HMM.

Such an 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 intuition does not necessarily hold for spectral learning. In particular, the magnitudes of the excluded singular values (numbers $k+1$ to m) are not necessarily predictive of the error of the resulting model. Moreover, in contrast to the statistical consistency of full-rank spectral learning, low-rank spectral learning can produce poor results even with infinite data, and even in cases where accurate low-rank models exist.

Throughout the discussion that follows we will assume an infinite training set. Though unrealistic, this allows us to isolate the effects of learning a low-rank model from finite-sample convergence issues. (Whether finite-sample issues would compound the difficulties of low-rank spectral learning in an interesting way—or, perhaps, alleviate them—remains an interesting and open question.) However, we cannot ignore the implications of finite training sets in practice; the need to accurately estimate singular vectors with small corresponding singular values is of particular concern precisely because very large quantities of data are required to do so (Benaych-Georges & Nadakuditi, 2012). Indeed, existing finite sample bounds typically have a term that grows like $O(1/\sigma^4)$, where σ is the smallest nonzero singular value of P_{21} (or equivalent) (Hsu et al., 2012; Boots et al., 2010a; Foster et al., 2012).

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.) For simplicity, we let

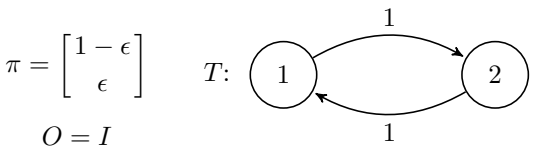

Figure 2. A modified two-state HMM.

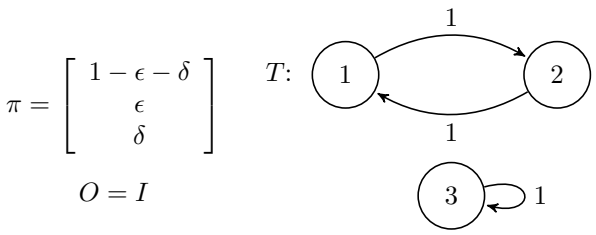

Figure 3. A three-state example HMM.

O be the identity, though this is not fundamental. 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

$$P_{21} = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}, \tag{5}$$

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 sign that reducing the rank will 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

$$P_{21} = \begin{bmatrix} 0 & \epsilon \\ 1-\epsilon & 0 \end{bmatrix}, \tag{6}$$

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. And yet, it 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. The conditions are assumptions about the world itself, and we cannot simply wish them stronger; if they do not hold, the entire learning procedure may fail.

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 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) given the rank constraint. 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

$$P_{21} = \begin{bmatrix} 0 & \epsilon & 0 \\ 1-\epsilon-\delta & 0 & 0 \\ 0 & 0 & \delta \end{bmatrix}, \tag{7}$$

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,

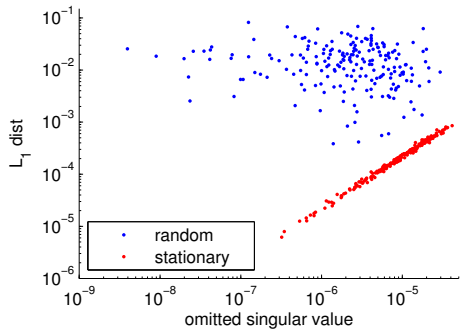

Figure 4. The relationship between variational distance and omitted singular values. Random HMMs are in blue (upper cloud) and HMMs initialized at the stationary distribution are in red.

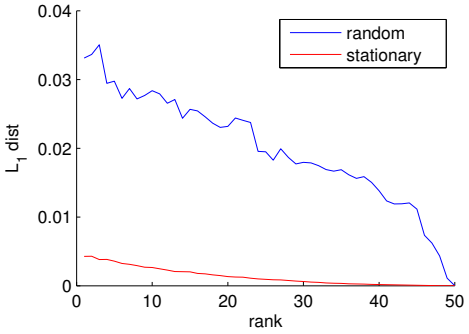

Figure 5. The relationship between variational distance and model rank. In blue (upper curve), π is random. In red, initialization occurs at the stationary distribution.

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.

4. The importance of π

Is there any good news? One apparent common thread in the examples of Section 3 is that they all rely on small values in the entries of π. This turns out to not be fundamental; for instance, the HMM given by $\pi = [0.25\ 0.25\ 0.25\ 0.25]^\top$ and

$$T = \begin{bmatrix} 0 & 1 & 0.5-\epsilon & 0 \\ 1 & 0 & 0.5-\epsilon & 1-\delta \\ 0 & 0 & 2\epsilon & 0 \\ 0 & 0 & 0 & \delta \end{bmatrix} \quad O = \begin{bmatrix} 1 & 0 & 0 & 1-\delta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \delta \end{bmatrix}$$

admits a near-perfect rank-three model, but rank-three spectral learning produces large errors when $\delta \gg \epsilon$.

However, this leads to a more general idea, which is that the bias introduced by π might somehow interfere with the underlying long-term structure of the HMM. Indeed, if π is replaced with the stationary distribution—that is, $T\pi = \pi$—then the examples seen so far become well-behaved. The first two yield poor-quality models since no good rank-one fit is possible, but the singular values now correctly indicate this fact. The remaining examples produce accurate models regardless of ϵ and δ. We conjecture that this may be true in a more general way, but leave details to future work.

Figure 4 contains a scatter plot in which each point represents a synthetic 10-state, 20-observation HMM for which a rank-9 spectral model has been learned. On the y-axis is the resulting L_1 variational distance at $t = 3$ (other choices of t were qualitatively similar), and on the x-axis is the magnitude of the omitted tenth singular value. Blue points correspond to HMMs generated randomly; entries of π, T, and O are sampled i.i.d. from the uniform distribution on $[0, 1]$, and then normalized to satisfy stochasticity. For red points, T and O are generated as before, but π is set to be the stationary distribution. This change creates an obvious and dramatic tightening of the relationship between singular values and variational error; moreover, the errors obtained by these models are uniformly lower.

Figure 5 shows the L_1 variational distance at $t = 3$ measured for a single 50-state, 100-observation HMM, generated randomly as before, across a range of spectral learning ranks. In blue, the original (random) initial distribution is used, while in red, π is replaced with the stationary distribution. The latter setting not only achieves lower error at all ranks (though the underlying model has changed as well), but reaches a point of diminishing returns more quickly. Note that the 40th singular value of P_{21} for this model is less than 10^{-6}, and yet, with random π, spectral learning still has not achieved error below 0.01 even at rank 44 out of 50.

Of course, we cannot in general change π since it is fixed by the world. Thus, for problems that fundamentally involve restarts to a fixed distribution, this result may be of limited value. However, for many real-world problems we observe a small number of long observation sequences, making the stationary distribution a natural measure. This is particularly common for predictive state representations, which are often used to model an agent's experience over long periods of time.

More generally, our results imply that the relationship between the initial distribution (which appears in some form in all spectral methods) and the remaining model parameters may be especially important for low-rank spectral learning.

References

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

Balle, Borja and Mohri, Mehryar. Spectral learning of general weighted automata via constrained matrix completion. In *Advances in Neural Information Processing Systems 25*, pp. 2168–2176, 2012.

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

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

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

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

Boots, Byron, Siddiqi, Sajid M, and Gordon, Geoffrey J. 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*, pp. 1369–1370. International Foundation for Autonomous Agents and Multiagent Systems, 2010b.

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

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

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

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

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

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