A Sober Look at Spectral Learning

Han Zhao and Pascal Poupart
{han.zhao, ppoupart}@uwaterloo.ca
David R. Cheriton School of Computer Science, University of Waterloo

Introduction

We do an empirical evaluation of spectral learning (SL) and expectation maximization (EM) for Hidden Markov Models (HMM) that reveals a gap between theory and practice:

Theory:

- SL is consistent and admits finite sample bounds.
- EM subject to local optima problem due to non-concave likelihoods.

Practice:

- SL suffers from negative probabilities.
- EM rarely gets stuck in local optima.
- EM often outperforms SL.

Our contribution: empirical evaluation

- The choice of rank hyperparameter and the amount of training data greatly influences SL.
- Likelihood function becomes unimodal with increased size of training data.

Background

A Hidden Markov Model $\mathcal{H} = \langle T, O, \pi \rangle$:

- Transition matrix $T \in \mathbb{R}^{m \times m}$, $T_{ij} = \Pr(s_{t+1} = i | s_t = j)$.
- Observation matrix $O \in \mathbb{R}^{n \times m}$, $O_{ij} = \Pr(o_t = i | s_t = j)$.
- Initial distribution $\pi \in \mathbb{R}^m$, $\pi_i = \Pr(s_1 = i)$.

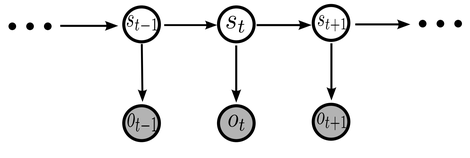


Figure 1: Graphical model of HMM.

Observable operators:
$A_x \triangleq T\text{diag}(O_{x,1}, \ldots, O_{x,m}), \forall x \in [n]$.
Equivalent parametrization: $\mathcal{H} = \langle \pi, A_x \rangle, \forall x \in [n]$.
Joint probability: $\Pr(x_{1:t}) = \mathbf{1}^T A_{x_t} \cdots A_{x_1} \pi$.

Spectral Learning (LearnHMM)

First three order moments:

- P_1, where $P_1[i] = \Pr(x_1 = i)$.
- $P_{2,1}$, where $P_{2,1}[ij] = \Pr(x_2 = i, x_1 = j)$.
- $P_{3,x,1}$, $\forall x \in [n]$, where $P_{3,x,1}[ij] = \Pr(x_3 = i, x_2 = x, x_1 = j)$.

Compute the following operators:

- $b_1 = U^T P_1$
- $b_\infty^T = P_1^T (U^T P_{2,1})^+$
- $B_x = U^T P_{3,x,1} (U^T P_{2,1})^+, \forall x \in [n]$

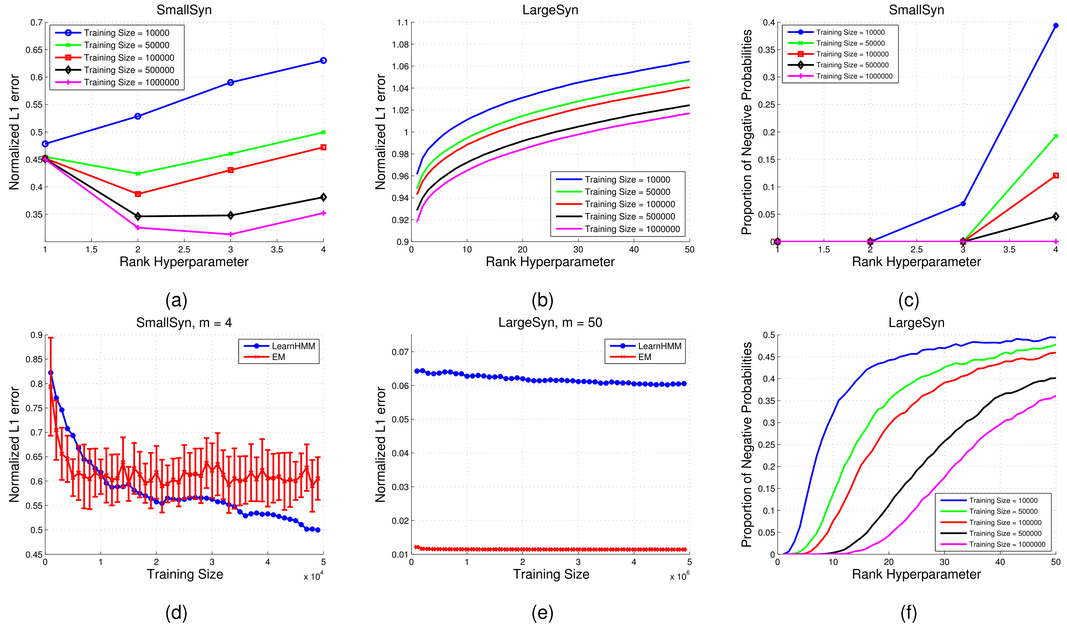
where U is the thin left singular matrix of $P_{2,1}$ such that $U^T O$ is invertible.
Joint probability: $\Pr(x_{1:t}) = b_\infty^T B_{x_t} \cdots B_{x_1} b_1$.
LearnHMM is shown by [1] to be consistent in joint probability and admits finite sample bounds.

For LearnHMM and EM, we report the normalized L_1 error of test sequence:

$$L_1 = \sum_{x_{1:t} \in \mathcal{T}} |\Pr(x_{1:t}) - \tilde{\Pr}(x_{1:t})|^{1/t} \quad (1)$$

For LearnHMM, we also report the proportion of negative probabilities in test dataset.

$$\text{NEG\_PROP} = \frac{|\{\tilde{\Pr}(x_{1:t}) < 0 \mid x_{1:t} \in \mathcal{T}\}|}{|\mathcal{T}|} \quad (2)$$



Experiment

Two synthetic experiments to compare LearnHMM and EM:

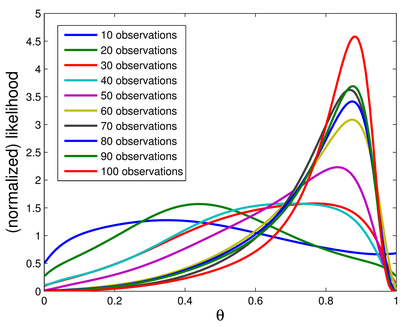
Table 1: Synthetic datasets

	SmallSyn	LargeSyn
# hidden states	4	50
# observations	8	100
test set size	4096	10,000
length of test sequence	4	50

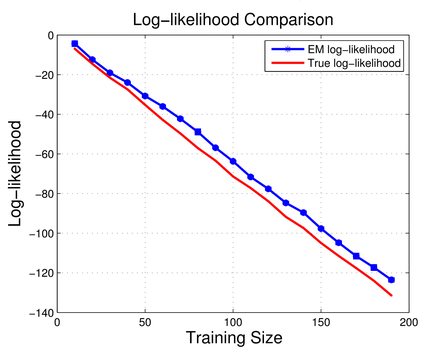
Each column vector of transition matrix T and observation matrix O is drawn from a uniform Dirichlet distribution. The stationary distribution of T is used as the initial distribution of the corresponding HMM to generate synthetic observation sequences.

Local optima of EM

The likelihood function of a single-parameter HMM with increased number of observations: $\mathcal{H} = \langle T = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}, O = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}, \pi = (0.5, 0.5) \rangle$



HMM with 2 hidden states, 2 observations and 4 parameters, EM v.s. true model parameter:



References

[1] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning Hidden Markov Models," *Journal of Computer and System Sciences*, vol. 78, pp. 1460–1480, Sept. 2012.

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School of Computer Science