Lecture 3: Spectral Learning of HMMs

Karl Stratos

October 8, 2018

The Problem

- We observe symbol sequences $\boldsymbol{x} \in [n]^*$ and their probabilities $p(\boldsymbol{x})$.
- God says there is some HMM (π, t, o) with m states such that

$$p(x_1 \ldots x_L) = \sum_{h_1 \ldots h_L \in [m]^L} \pi(h_1) o(x_1|h_1) \prod_{l=2}^{L} t(h_t|h_{t-1}) o(x_t|h_t)$$

- **Goal.** Learn $\hat{p} : [n]^* \to [0, 1]$ satisfying

$$\hat{p}(\boldsymbol{x}) = p(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in [n]^*$$

Two Approaches to Spectral Learning of HMMs

- Special case of learning weighted finite automata (Balle et al., 2014; Hsu et al., 2008)

- Dimensionality reduction followed by the method of moments (Foster et al., 2012)

Overview

- Spectral Learning of WFAs
- Dimensionality Reduction + Method of Moments

Weighted Finite Automaton (WFA)

- Hypothesis class of **WFA**s

$$\mathcal{H} := \left\{ (a_0, \{A^\sigma\}_{\sigma \in [n]^*}, a_\infty) : a_0, a_\infty \in \mathbb{R}^m, A^\sigma \in \mathbb{R}^{m \times m}, m \in \mathbb{N} \right\}$$

- $A \in \mathcal{H}$ induces $f_A : [n]^* \to \mathbb{R}$ by

$$f_A(\boldsymbol{x}) = a_0^\top \underbrace{A^{x_1} \cdots A^{x_L}}_{A^{\boldsymbol{x}}} a_\infty$$

- Given access to input-output pairs of $f : [n]^* \to \mathbb{R}$, find a **minimal WFA** computing f

$$A_f \in \operatorname*{arg\,min}_{A \in \mathcal{H}:\ f = f_A} m_A$$

Hankel Matrix

- **Theorem** (Carlyle and Paz, 1971). Define $H_f \in \mathbb{R}^{\infty \times \infty}$ by

  $$[H_f]_{\boldsymbol{y}\boldsymbol{z}} := f(\boldsymbol{y}\boldsymbol{z}) \qquad\qquad \forall \boldsymbol{y}, \boldsymbol{z} \in [n]^*$$

  (called **Hankel matrix** associated with f). Then

  $$\operatorname{rank}(H_f) = \min_{A \in \mathcal{H}:\ f = f_A} m_A$$

- Thus if $B \in \mathcal{H}$ satisfies $f = f_B$ and $m_B = \operatorname{rank}(H_f)$, then B is a minimal WFA computing f.
- A **sufficient Hankel sub-block** is $\widetilde{H}_f \in \mathbb{R}^{|\mathcal{P}| \times |\mathcal{S}|}$ indexed by some finite $\mathcal{P}, \mathcal{S} \subset [n]^*$ such that $\epsilon \in \mathcal{P} \cap \mathcal{S}$ and

  $$\operatorname{rank}\left(\widetilde{H}_f\right) = \operatorname{rank}(H_f)$$

Derivation of a Spectral Algorithm

- Consider any $f_A : [n]^* \to \mathbb{R}$ where $m_A = m$.
- Since $[\widetilde{H}_f]_{\boldsymbol{y}\boldsymbol{z}} = a_0^\top A^{\boldsymbol{y}} A^{\boldsymbol{z}} a_\infty$, a sufficient Hankel sub-block admits a natural rank-m decomposition

$$\underbrace{\widetilde{H}_f}_{|\mathcal{P}|\times|\mathcal{S}|} = \underbrace{P}_{|\mathcal{P}|\times m} \underbrace{S}_{m\times|\mathcal{S}|} \qquad [P]_{\boldsymbol{y},:} := a_0^\top A^{\boldsymbol{y}}, \; [S]_{:,\boldsymbol{z}} := A^{\boldsymbol{z}} a_\infty$$

- If we define $[\widetilde{H}_f^x]_{\boldsymbol{y}\boldsymbol{z}} := f(\boldsymbol{y}x\boldsymbol{z})$ for $x \in [n]$, similarly we have

$$\underbrace{\widetilde{H}_f^x}_{|\mathcal{P}|\times|\mathcal{S}|} = \underbrace{P}_{|\mathcal{P}|\times m} \underbrace{A^x}_{m\times m} \underbrace{S}_{m\times|\mathcal{S}|}$$

- Thus if God gives us P and S, we can recover A by

$$A^x = P^+ \widetilde{H}_f^x S^+ \qquad a_0^\top = [P]_{\epsilon,:} \qquad a_\infty = [S]_{:,\epsilon}$$

Derivation of a Spectral Algorithm (Cont.)

- Consider *any* rank-m decomposition

$$\underbrace{\widetilde{H}_f}_{|\mathcal{P}|\times|\mathcal{S}|} = \underbrace{U}_{|\mathcal{P}|\times m}\underbrace{W}_{m\times|\mathcal{S}|}$$

- **Claim.** $B \in \mathcal{H}$ defined by

$$B^x = U^+ \widetilde{H}_f^x W^+ \qquad b_0^\top = [U]_{\epsilon,:} \qquad b_\infty = [W]_{:,\epsilon}$$

  is a minimal WFA computing f_A.

- **Proof.** Follows from the fact that

$$B^x = G A^x G^{-1} \qquad b_0^\top = a_0^\top G^{-1} \qquad b_\infty = G a_\infty$$

  where $G := U^+ P$ with inverse $G^{-1} = S W^+$.

Application to HMM Learning

- Organizie HMM parameters as vector/matrices (assumed to be full-rank):

$$\pi \in [0,1]^{m} \qquad [\pi]_h = \pi(h)$$
$$T \in [0,1]^{m\times m} \qquad [T]_{:h} = t(\cdot|h)$$
$$O \in [0,1]^{n\times m} \qquad [O]_{:h} = o(\cdot|h)$$

- **Matrix form of the forward algorithm**

$$p(x_1 \ldots x_L) = \pi^\top \underbrace{\mathrm{diag}(O^\top \delta_{x_1})T}_{A^{x_1}} \cdots \underbrace{\mathrm{diag}(O^\top \delta_{x_L})T}_{A^{x_L}} 1$$

- Sufficient Hankel sub-block $P_{1,2} \in [0,1]^{(n+1)\times(n+1)}$ given by

$$[P_{1,2}]_{yz} := p(yz) \qquad \forall y,z \in [n] \cup \{\epsilon\}$$

(Exercise: to show this, express $P_{1,2}$ in terms of π, T, O.)

Algorithm

1. Estimate $\widehat{P}_{1,2}, \widehat{P}_{1,x,3} \in [0,1]^{(n+1)\times(n+1)}$ from HMM samples:

$$[\widehat{P}_{1,2}]_{yz} \approx p(yz) \qquad [\widehat{P}_{1,x,3}]_{yz} \approx p(yxz) \qquad \forall y,z \in [n] \cup \{\epsilon\}$$

2. Rank-m SVD

$$\widehat{P}_{1,2} \approx \underbrace{\widehat{U}}_{(n+1)\times m} \underbrace{\widehat{\Sigma}}_{m\times m} \underbrace{\widehat{V}^\top}_{m\times(n+1)}$$

3. Let $\widehat{W} = \widehat{\Sigma}\widehat{V}^\top$ and compute

$$\widehat{B}^x = \widehat{U}^\top \widehat{P}_{1,x,3} \widehat{W}^+ \qquad \hat{b}_0^\top = [\widehat{U}]_{\epsilon,:} \qquad \hat{b}_\infty = [\widehat{W}]_{:,\epsilon}$$

4. Given any $x_1 \ldots x_L \in [n]^*$, predict

$$\hat{p}(x_1 \ldots x_L) = \hat{b}_0^\top \widehat{B}^{x_1} \cdots \widehat{B}^{x_L} \hat{b}_\infty$$

Overview

- Spectral Learning of WFAs
- Dimensionality Reduction + Method of Moments

Idea

- Let $U \in \mathbb{R}^{n \times m}$ be any matrix such that $U^\top O$ is invertible.
- Calculate m-dimensional representation of first three observations $x_1, x_2, x_3 \in [n]$ under HMM by

$$y_i = U^\top \delta_{x_i}$$

- Verify that

$$\begin{aligned}
\mu := \mathbf{E}\left[y_1\right] \quad &= U^\top O \pi \\
\Sigma := \mathbf{E}\left[y_1 y_2^\top\right] \quad &= U^\top O \mathrm{diag}(\pi) T O^\top U \\
K^x := \mathbf{E}\left[[[x_2 = x]]\, y_1 y_3^\top\right] \quad &= U^\top O \mathrm{diag}(\pi) T \mathrm{diag}(O^\top \delta_x) T O^\top U
\end{aligned}$$

Idea (Cont.)

- Thus if we define

$$\begin{aligned} c_0^\top &:= \mu^\top && = \pi^\top (O^\top U) \\ c_\infty &:= \Sigma^{-1}\mu && = (O^\top U)^{-1} 1 \\ C^x &:= \Sigma^{-1} K^x && = (O^\top U)^{-1} \mathsf{diag}(O^\top \delta_x) T (O^\top U) \end{aligned}$$

it follows that

$$p(x_1 \ldots x_L) = c_0^\top C^{x_1} \cdots C^{x_L} c_\infty$$

How to Choose U

- What $U \in \mathbb{R}^{n \times m}$ (such that $U^\top O$ is invertible) should we use?
  - Assume $|U_{i,j}| \leq 1$.

- Answer: whatever U that makes estimation $\hat{\theta}$ easier

- Challenge in analysis: we need to estimate the matrix *inverse*

$$\Sigma^{-1}$$

by first estimating Σ and then taking the inverse of *that* estimate:

$$\widehat{\Sigma}^{-1}$$

First Lemma

Given N samples of y_1, y_2 to estimate $\Sigma = \mathbf{E}\left[y_1 y_2^\top\right]$,

$$\Pr\left(\left|\left|\widehat{\Sigma} - \Sigma\right|\right|_2 \leq \underbrace{m\sqrt{\frac{\ln\frac{2m}{\delta}}{N}}}_{J}\right) \geq 1 - \delta$$

Proof

$$
\begin{aligned}
\Pr\left(\left\|\widehat{\Sigma} - \Sigma\right\|_2 \geq \epsilon\right) &\leq \Pr\left(m \left\|\widehat{\Sigma} - \Sigma\right\|_{\max} \geq \epsilon\right) \\
&\leq \sum_{i,j=1}^{m} \Pr\left(\left|\widehat{\Sigma}_{i,j} - \Sigma_{i,j}\right| \geq \frac{\epsilon}{m}\right) \\
&\leq 2m^2 \exp\left(-2N \frac{\epsilon^2}{m^2}\right) \\
&= \delta
\end{aligned}
$$

holds if

$$
\epsilon = m \sqrt{\frac{\ln \frac{2m}{\delta}}{N}}
$$

Second Lemma

Assuming $N \geq \frac{16J^2}{\sigma_m(\Sigma)^2}$,

$$\Pr\left(\left|\left|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right|\right|_{\max} \leq \frac{4J}{\sigma_m(\Sigma)^2}\right) \geq 1 - \delta$$

Key matrix perturbation tools:

$$\left|\left|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right|\right|_2 \leq 2\max\left\{\left|\left|\widehat{\Sigma}^{-1}\right|\right|_2^2, \left|\left|\Sigma^{-1}\right|\right|_2^2\right\}\left|\left|\widehat{\Sigma} - \Sigma\right|\right|_2$$

$$|\hat{\sigma}_i - \sigma_i| \leq \left|\left|\widehat{\Sigma} - \Sigma\right|\right|_2 \qquad \forall i \in [m]$$

Proof

Using $\sigma_m - \hat{\sigma}_m \leq J$ (w.p. $1-\delta$),

$$\frac{1}{\hat{\sigma}_m} \leq \frac{1}{\sigma_m - J}$$

If $N \geq \frac{16J^2}{\sigma_m^2}$, then $\sigma_m \geq 4J$ so $\sigma_m - J \geq \frac{3\sigma_m}{4}$ and

$$\left(\frac{1}{\hat{\sigma}_m - J}\right)^2 \leq \left(\frac{4}{3\sigma_m}\right)^2 \leq \frac{2}{\sigma_m^2}$$

It follows that

$$\begin{aligned} \max\left\{\left|\left|\widehat{\Sigma}^{-1}\right|\right|_2^2, \left|\left|\Sigma^{-1}\right|\right|_2^2\right\} &= \max\left\{\left(\frac{1}{\sigma_m}\right)^2, \left(\frac{1}{\hat{\sigma}_m}\right)^2\right\} \\ &\leq \left(\frac{1}{\hat{\sigma}_m - J}\right)^2 \leq \frac{2}{\sigma_m^2} \end{aligned}$$

Proof (Cont.)

From previous two slides and the first lemma,

$$\Pr\left(\left\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right\|_2 \geq \frac{4J}{\sigma_m^2}\right) \leq \delta$$

Thus

$$\Pr\left(\left\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right\|_{\max} \geq \frac{4J}{\sigma_m^2}\right) \leq \Pr\left(\left\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\right\|_2 \geq \frac{4J}{\sigma_m^2}\right) \leq \delta$$

Sample Complexity

$$
\begin{aligned}
\left|\hat{\theta} - \theta\right| \leq \frac{4J}{\sigma_m(\Sigma)^2} \quad &\Rightarrow \quad \theta - \frac{4J}{\sigma_m(\Sigma)^2} \leq \hat{\theta} \leq \theta - \frac{4J}{\sigma_m(\Sigma)^2} \\
&\Rightarrow \quad 1 - \frac{4J}{\sigma_m(\Sigma)^2\theta} \leq \frac{\hat{\theta}}{\theta} \leq 1 - \frac{4J}{\sigma_m(\Sigma)^2\theta} \\
&\Rightarrow \quad 1 - \frac{4J}{\sigma_m(\Sigma)^2\Lambda} \leq \frac{\hat{\theta}}{\theta} \leq 1 - \frac{4J}{\sigma_m(\Sigma)^2\Lambda} \\
&\Rightarrow \quad \left(1 - \frac{4J}{\sigma_m(\Sigma)^2\Lambda}\right)^{2L+3} \leq \frac{\hat{p}}{p} \leq \left(1 - \frac{4J}{\sigma_m(\Sigma)^2\Lambda}\right)^{2L+3} \\
&\Rightarrow \quad 1 - \epsilon \leq \frac{\hat{p}}{p} \leq 1 + \epsilon
\end{aligned}
$$

holds w.p. at least $1 - \delta$ when

$$
N = O\left(\frac{m^2 \ln \frac{m}{\delta}}{((1+\epsilon)^{1/(2L+3)} - 1)^2 \sigma_m(\Sigma)^4 \Lambda^2}\right)
$$

So Which U?

- Choose $U \in \mathbb{R}^{n \times m}$ so that

$$\sigma_m\left(\Sigma\right) = \sigma_m\left(\mathbf{E}\left[U^\top \delta_{x_1} \delta_{x_2}^\top U\right]\right) = \sigma_m\left(U^\top P_{1,2} U\right)$$

  is large!

- In particular, if U is the top m left singular vectors of $P_{1,2} \in \mathbb{R}^{n \times n}$,

$$\sigma_m\left(\Sigma\right) = \sigma_m\left(P_{1,2}\right)$$