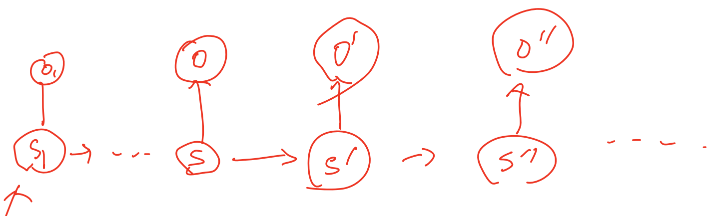
Spectral learning for HMMs

① HMM



$s_1 \sim \mu$ $\quad o \sim O(\cdot|s) \quad s' \sim T(\cdot|s),\ o' \sim O(\cdot|s')$

Q1: Given $o_1 o_2 \cdots o_T$, $\Pr(o_1, o_2, \cdots o_T)$

Q2: Given history $o_1 \cdots o_t$ $\Pr(o_{t+1}, o_{t+2} \mid o_1 \cdots o_t)$

$$\Pr(o_1 o_2 \cdots o_t)$$

$$= \sum_{s_1 s_2 \cdots s_t} \Pr(s_1, o_1, s_2\, o_2 \cdots s_t, o_t)$$

$$= \sum_{s_1 s_2 \cdots s_t} \mu(s_1) \cdot O(o_1|s_1) T(s_2|s_1) O(o_2|s_2) \cdots T(s_t|s_{t-1}) \cdot O(o_t|s_t)$$

learning: T & O are unknown

$$D = \Big\{ \underbrace{\{o_1, o_2, o_3 \cdots o_T\}}_{\tau_n}{}_{n=1}^{N} \Big\}$$

MLE:

$$\max_{T,O} \ln\left(\prod_{n=1}^{N} \Pr(\tau_n; T, O)\right)$$

$$= \max_{TO} \sum_{n=1}^{N} \ln \Pr(\tau_n; TO) = \max_{TO} \sum_{n=1}^{N} \ln \Big[\sum_{s_1 \cdots s_t} \mu(s_1) O(o_1|s_1) T(s_2|s_1) \cdots T(s_{t-1}|s_t) \Big]$$

$\cdot O(o_t|s_t) \,|$

Spectral learning

$s \in [n], \quad o \in [m]$

observable operator $A_o \in \mathbb{R}^{n \times n}$

Definition: $A_o = \underbrace{T}_{\in \mathbb{R}^{n\times n}} \cdot \underbrace{\mathrm{Diag}\left[O(o|1), O(o|2) \cdots, O(o|n) \right]}_{\mathbb{R}^{n\times n}}$

$$A_o[i,j] = T(i|j) \cdot O(o|j)$$
$$= \Pr(s' = i, o \mid s = j)$$

Lemma:

$$\Pr(o_1 \cdots o_t) = \mathbf{1}^T A_{o_t} A_{o_{t-1}} \cdots A_{o_1} \cdot \mu$$

initial latent Distribution

Proof:

$$\Pr(s_2, o_1) = \sum_{s_1 \in [n]} \overbrace{\mu(s_1) \underbrace{O(o_1|s_1) T(s_2|s_1)}_{A_{o_1}[s_2, s_1]}}^{\Pr(o_1, s_1, s_2)}$$

$$= \mu^T A_{o_1}[s_2, :]^T$$

$$\Pr(s_3, o_2, o_1) = \sum_{s_2 \in [n]} P(s_2, o_1) P(s_3, o_2 | o_1, s_2)$$

$$= \sum_{s_2 \in [n]} \Pr(s_2, o_1) \underbrace{P(s_3, o_2 | s_2)}_{A_{o_2}[s_3, s_2]}$$

$$= \sum_{s_2 \in [n]} \left(\mu^T A_{o_1}[s_2,:]^T \right) \cdot \left(A_{o_2}[s_3, s_2] \right)$$

$$= \left(\mu^T A_{o_1} \right) \left(A_{o_2}[s_3,:] \right)^T$$

$$Pr(o_1 \cdots o_t) = 1^T A_{o_t} A_{o_{t-1}} \cdots A_{o_1} \cdot \mu \qquad C \in \mathbb{R}^{n\times n}, \; C^{-1}$$

$$= \underbrace{1^T C^{-1}}_{b_\infty^T} \underbrace{C A_{o_t} C^{-1}}_{B_{o_t}} C A_{o_{t-1}} \cdots C^{-1} \underbrace{C A_{o_1} C^{-1}}_{B_{o_1}} \underbrace{C \mu}_{b_1}$$

Assumption: $\underset{\#\text{ of } o}{m} \geq \underset{\#\text{ of } s}{n}$, T & O are full Rank

observability:

$p \in \Delta(n)$ $(s \sim p)$ $\qquad q := O p$ $(o \sim q)$ $\qquad O^+ O = I$

$$\Rightarrow p = O^+ q$$

$P_1, \; P_{2,1}, \; P_{3,o,1}, \; o \in [m]$

$P_1[i] = P(o_1 = i)$; $\quad P_{2,1}[i,j] = P(o_2 = i, o_1 = j)$

$$\approx \frac{\sum_{i=1}^{N} 1(o_1^i = j, o_2^i = i)}{N}$$

$\forall o \in [m]$

$$P_{3,o,1}[i,j] = Pr(o_3 = i, o_2 = o, o_1 = j)$$

$$\approx \sum_{i=1}^{N} 1(o_1^i = j, o_2^i = o, o_3^i = i) / N$$

$O[i,j] = O(o = i \mid s = j)$

$C = U^T O$, U is the left singular vectors of $P_{2,1}$

$P_{2,1} = U \Sigma V^T$

lemma: $U^T O$ is full Rank

Proof: $P_{2,1}[i,j] = P(o_2 = i, o_1 = j) = \sum_{s_1 s_2} \mu(s_1) \cdot O(j|s_1) T(s_2|s_1) O(i|s_2)$

$P_{2,1} = O T \operatorname{diag}(\mu) O^T$

$P_{2,1}[i,j] = O[i,:] T \operatorname{Diag}(\mu) (O^T)[:,j]$

$\operatorname{Span}(U) = \operatorname{Span}(P_{2,1}) \subseteq \operatorname{span}(O)$

Since O is full column Rank: $O^T (O^T)^+ = I$

$P_{2,1} (O^T)^+ = O T \operatorname{diag}(\mu)$

$\Rightarrow P_{2,1} (O^T)^+ (\operatorname{Diag}(\mu))^{-1} = O T$

$\Rightarrow P_{2,1} (O^T)^+ (\operatorname{Diag}(\mu))^{-1} T^{-1} = O$

$\Rightarrow \operatorname{span}(O) \subseteq \operatorname{span}(P_{2,1}) = \operatorname{span}(U)$

$\Rightarrow \operatorname{span}(U) = \operatorname{span}(O)$

$\Rightarrow (U^T O)$ is full Rank