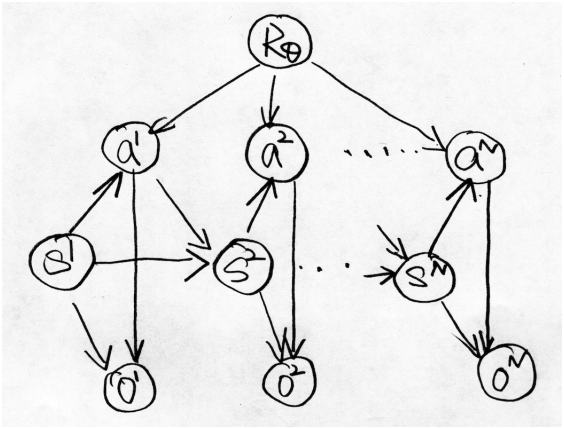

(a)

MAP$(R_\theta \mid o^1, o^2, ..., o^N)$

$$\arg\max_{R_\theta} Pr(R_\theta \mid o^1, o^2 ... o^N) \tag{1}$$

$$= \arg\max_{R_\theta} Pr(o^1, o^2 ... o^N \mid R_\theta) \cdot Pr(R_\theta) \tag{2}$$

$$= \arg\max_{R_\theta} (\log Pr(o^1, o^2 ... o^N \mid R_\theta) \cdot Pr(R_\theta)) \tag{3}$$

$$= \arg\max_{R_\theta} (\log Pr(o^1, o^2 ... o^N \mid R_\theta) + \log Pr(R_\theta)) \tag{4}$$

$$= \arg\max_{R_\theta} (\log \sum_{s^1, a^1, ..., s^N, a^N} Pr(o^1, o^2 ..., o^N, s^1, a^1 ..., s^N, a^N \mid R_\theta) + \log Pr(R_\theta)) \tag{5}$$

Denote: $\tau, L_\theta^{lh}, L_\theta^{pr}$

$$\tau = \{(s^1, a^1), (s^2, a^2), ..., (s^N, a^N)\} \tag{6}$$

$$L_\theta^{pr} = \log Pr(R_\theta) \tag{7}$$

$$L_\theta^{lh} = \log \sum_\tau Pr(o^1, o^2 ..., \tau \mid R_\theta) \tag{8}$$

Goal:

$$L_\theta = L_\theta^{pr} + L_\theta^{lh} \tag{9}$$

$$\frac{\partial L_\theta}{\partial \theta} = \frac{\partial L_\theta^{pr}}{\partial \theta} + \frac{\partial L_\theta^{lh}}{\partial \theta} \tag{10}$$

(b)

Derivative of the log prior: L_θ^{pr}

$$L_\theta^{pr} = \log Pr(R_\theta) \tag{11}$$

$$Pr(R_\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma_{R_\theta}|^{1/2}} \exp\left(-\frac{1}{2}(R_\theta - \mu_{R_\theta})^T \Sigma_{R_\theta}^{-1}(R_\theta - \mu_{R_\theta})\right) \tag{12}$$

$$\frac{\partial log Pr(R_\theta)}{\partial \theta} = \frac{1}{Pr(R_\theta)} \cdot \frac{\partial Pr(R_\theta)}{\partial R_\theta} \cdot \frac{\partial R_\theta}{\partial \theta} \tag{13}$$

$$= \frac{1}{Pr(R_\theta)} \cdot (-Pr(R_\theta) \cdot \Sigma_{R_\theta}^{-1}(R_\theta - \mu_{R_\theta})) \cdot \frac{\partial R_\theta}{\partial \theta} \tag{14}$$

$$= -\Sigma_{R_\theta}^{-1}(R_\theta - \mu_{R_\theta}) \cdot \frac{\partial R_\theta}{\partial \theta} \tag{15}$$

(c)

Derivative of the log likelihood: L_θ^{lh}

$$L_\theta^{lh} = \log \sum_\tau Pr(o^1, o^2 ..., \tau \mid R_\theta) \tag{16}$$

$$= \log \sum_\tau Pr(s^1) \prod_{i=1}^{N} Pr(o^i \mid s^i, a^i) \prod_{j=2}^{N} Pr(s^j \mid s^{j-1}, a^{j-1}) \prod_{z=1}^{N} Pr(a^z \mid s^z, R_\theta) \tag{17}$$

Denote: $h_\theta(\tau), \pi_\theta(a^z \mid s^z)$

$$\pi_\theta(a^z \mid s^z) = Pr(a^z \mid s^z, R_\theta) \tag{18}$$

$$h_\theta(\tau) = Pr(s^1) \prod_{i=1}^{N} Pr(o^i \mid s^i, a^i) \prod_{j=2}^{N} Pr(s^j \mid s^{j-1}, a^{j-1}) \prod_{z=1}^{N} \pi_\theta(a^z \mid s^z) \tag{19}$$

$$L_\theta^{lh} = \log \sum_\tau h_\theta(\tau) \tag{20}$$

$$\frac{\partial L_\theta^{lh}}{\partial \theta} = \frac{1}{\sum_\tau h_\theta(\tau)} \cdot \frac{\partial \sum_\tau h_\theta(\tau)}{\theta} \tag{21}$$

$$= \frac{1}{\sum_\tau h_\theta(\tau)} \cdot \sum_\tau \frac{\partial h_\theta(\tau)}{\partial \theta} \tag{22}$$

(c)

Derivative of $h_\theta(\tau)$

$$h_\theta(\tau) = Pr(s^1) \prod_{i=1}^{N} Pr(o^i \mid s^i, a^i) \prod_{j=2}^{N} Pr(s^j \mid s^{j-1}, a^{j-1}) \prod_{z=1}^{N} \pi_\theta(a^z \mid s^z) \tag{23}$$

$$= c(\tau) \prod_{z=1}^{N} \pi_\theta(a^z \mid s^z) \tag{24}$$

$$\frac{\partial h_\theta(\tau)}{\theta} = c(\tau) \cdot (\sum_{z=1}^{N} (\frac{\partial \pi_\theta(a^z \mid s^z)}{\partial \theta} \prod_{j \neq z}^{N} \pi_\theta(a^j \mid s^j))) \tag{25}$$

$$\pi_\theta(a^z \mid s^z) = \frac{\exp(Q_\theta^*(s^z, a^z))}{\sum_{a' \in A} \exp(Q_\theta^*(s^z, a'))} \tag{26}$$

$$\frac{\partial \pi_\theta(a^z \mid s^z)}{\partial \theta} = \pi_\theta(a^z \mid s^z)(\frac{\partial Q_\theta^*(s^z, a^z)}{\partial \theta} - \sum_{a' \in A} \pi_\theta(a' \mid s^z) \frac{\partial Q_\theta^*(s^z, a')}{\partial \theta}) \tag{27}$$

(d)

Derivative of $Q_\theta^*(s, a)$

$$\frac{\partial Q_\theta^*(s,a)}{\partial \theta} = \phi_\theta(s,a) \tag{28}$$

Solving fixed point equation:

$$\phi_\theta(s,a) = \frac{\partial R_\theta(s,a)}{\partial \theta} + \gamma \sum_{s' \in S} T(s,a,s') \sum_{a' \in A} \pi_\theta(a' \mid s') \phi_\theta(s',a') \tag{29}$$

(iter)

initialize θ_0
For $k = 1$ to $MaxIter$:
Compute $\frac{\partial L_\theta}{\partial \theta}$
$\theta_k = \theta_{k-1} + \sigma \frac{\partial L_\theta}{\partial \theta}$