Training Hidden Markov Models

Hao Tang

Automatic Speech Recognition—ASR Lecture 5
31 January 2022

Hidden Markov Models

- States S, e.g., $\{1, 2, 3\}$ ($J = |S|$)
- Prior probabilities π, e.g., $[1, 0, 0]$
- Transition probability $p(q' = j | q = i) = a_{ij}$

$$A = \begin{bmatrix} 0.1 & 0.9 & 0 \\ 0 & 0.4 & 0.6 \\ 0 & 0 & 1 \end{bmatrix}$$

- Emission probability $p(x | q = j) = b_j(x)$
- Observations $X = x_{1:T} = x_1, x_2, \ldots, x_T$

Statistical Queries

- $p(X, Q)$
- $\text{argmax}_Q\, p(Q|X)$
- $p(X)$
- $p(Q|X) = \frac{p(X,Q)}{p(X)}$

Joint Probability

$$p(X, Q) = p(q_1)p(x_1|q_1)\prod_{t=2}^{T} p(q_t|q_{t-1})p(x_t|q_t)$$

Viterbi Algorithm

$$\underset{Q}{\operatorname{argmax}}\, p(Q|X) = \underset{Q}{\operatorname{argmax}}\, p(X, Q)$$

$$V_{q_t}(t) = \max_{q_{t-1}} p(q_t|q_{t-1}) p(x_t|q_t) V_{q_{t-1}}(t-1)$$
$$V_{q_1}(1) = p(q_1)$$

$$V_j(t) = \max_{i=1,\dots,J} a_{ij} b_j(x_t) V_i(t-1)$$
$$V_j(1) = \pi_j$$

Forward Algorithm

$$\alpha_{q_t}(t) = \sum_{q_{1:t-1}} p(x_{1:t}, q_{1:t}) = p(x_{1:t}, q_t)$$

$$\alpha_{q_t}(t) = \sum_{q_{t-1}} p(q_t|q_{t-1}) p(x_t|q_t) \alpha_{q_{t-1}}(t-1)$$

$$\alpha_{q_1}(1) = p(q_1)$$

$$\alpha_j(t) = \sum_{i=1,\dots,J} a_{ij} b_j(x_t) \alpha_i(t-1)$$

$$\alpha_j(1) = \pi_j$$

Forward Algorithm

$$\alpha_{q_t}(t) = \sum_{q_{1:t-1}} p(x_{1:t}, q_{1:t}) = p(x_{1:t}, q_t)$$

$$\alpha_{q_t}(t) = \sum_{q_{t-1}} p(q_t|q_{t-1})p(x_t|q_t)\alpha_{q_{t-1}}(t-1)$$

$$\alpha_{q_1}(1) = p(q_1)$$

$$\alpha_j(t) = \sum_{i=1,\ldots,J} a_{ij} b_j(x_t)\alpha_i(t-1)$$

$$\alpha_j(1) = \pi_j$$

Backward Algorithm

$$\beta_{q_t}(t) = \sum_{q_{t+1:T}} p(x_{t+1:T}, q_{t+1:T} | q_t) = p(x_{t+1:T} | q_t)$$

$$\beta_{q_{t-1}}(t-1) = \sum_{q_t} p(q_t | q_{t-1}) p(x_t | q_t) \beta_{q_t}(t)$$

$$\beta_{q_T}(T) = 1$$

$$\beta_i(t-1) = \sum_{j=1}^{J} a_{ij} b_j(x_t) \beta_j(t)$$

$$\beta_i(T) = 1$$

Backward Algorithm

$$\beta_{q_t}(t) = \sum_{q_{t+1:T}} p(x_{t+1:T}, q_{t+1:T} | q_t) = p(x_{t+1:T} | q_t)$$

$$\beta_{q_{t-1}}(t-1) = \sum_{q_t} p(q_t | q_{t-1}) p(x_t | q_t) \beta_{q_t}(t)$$

$$\beta_{q_T}(T) = 1$$

$$\beta_i(t-1) = \sum_{j=1}^{J} a_{ij} b_j(x_t) \beta_j(t)$$

$$\beta_i(T) = 1$$

Training HMMs

- Parameters of an HMM: $\lambda = \{\pi, A, \text{parameters in } b_j(x)\}$
- Data: R i.i.d. utterances $X^1, X^2, \ldots, X^R$
- Likelihood of λ: $\prod_{r=1}^{R} p_\lambda(X^r)$
- Goal: find λ to maximize the likelihood $\prod_{r=1}^{R} p_\lambda(X^r)$

$$\operatorname*{argmax}_{\lambda} \prod_{r=1}^{R} p_\lambda(X^r)$$

- We will talk about three approaches!

Stochastic Gradient Descent

Repeat:

choose one of the R utterances X^r

$$\lambda^{s+1} \leftarrow \lambda^s - \nabla_\lambda p_\lambda(X^r)|_{\lambda=\lambda^s}$$

Computing the Gradient

- Backpropagation

Computing the Gradient

- Backpropagation
- Closed-form solution

$$p(X) = \sum_{Q} p(X, Q)$$

Computing the Gradient

- Backpropagation
- Closed-form solution

$$p(X) = \sum_{Q} p(X, Q) = \sum_{q_1} p(q_1) p(x_1 | q_1) p(x_{2:T} | q_1)$$

Computing the Gradient

- Backpropagation

- Closed-form solution

$$p(X) = \sum_{Q} p(X, Q) = \sum_{q_1} p(q_1) p(x_1 | q_1) p(x_{2:T} | q_1)$$

Computing the Gradient

- Backpropagation
- Closed-form solution

$$
\begin{aligned}
p(X) &= \sum_{Q} p(X, Q) = \sum_{q_1} p(q_1) p(x_1 | q_1) p(x_{2:T} | q_1) \\
&= \sum_{q_1} p(q_1) p(x_1 | q_1) \beta_{q_1}(1)
\end{aligned}
$$

Computing the Gradient

- Backpropagation

- Closed-form solution

$$p(X) = \sum_{Q} p(X, Q) = \sum_{q_1} p(q_1)p(x_1|q_1)p(x_{2:T}|q_1)$$
$$= \sum_{q_1} p(q_1)p(x_1|q_1)\beta_{q_1}(1)$$

$$\frac{\partial p(X)}{\partial p(q_1)} = p(x_1|q_1)\beta_{q_1}(1)$$

Computing the Gradient

- Backpropagation

- Closed-form solution

$$p(X) = \sum_{Q} p(X, Q) = \sum_{q_1} p(q_1)p(x_1|q_1)p(x_{2:T}|q_1)$$
$$= \sum_{q_1} p(q_1)p(x_1|q_1)\beta_{q_1}(1)$$

$$\frac{\partial p(X)}{\partial p(q_1)} = p(x_1|q_1)\beta_{q_1}(1) \qquad \frac{\partial p(X)}{\partial \pi_i} = b_i(x_1)\beta_i(1)$$

- What would be the best parameters had we know the hidden sequence?

- What would be the best parameters had we know the hidden sequence?

$$\sum_{r=1}^{R} \log p(X^r, Q^r)$$
$$= \sum_{r=1}^{R} \log \left[p(q_1^r) p(x_1^r | q_1^r) \prod_{t=2}^{T_r} p(q_t^r | q_{t-1}^r) p(x_t^r | q_t^r) \right]$$

- What would be the best parameters had we know the hidden sequence?

$$\sum_{r=1}^{R} \log p(X^r, Q^r)$$

$$= \sum_{r=1}^{R} \log \left[p(q_1^r) p(x_1^r | q_1^r) \prod_{t=2}^{T_r} p(q_t^r | q_{t-1}^r) p(x_t^r | q_t^r) \right]$$

$$= \sum_{r=1}^{R} \left[\log \pi_{q_1^r} + \log b_{q_1^r}(x_1^r) + \sum_{t=2}^{T_r} \left(\log a_{q_t^r q_{t-1}^r} + \log b_{q_t^r}(x_t^r) \right) \right]$$

- What would be the best parameters had we know the hidden sequence?

$$
\begin{aligned}
&\sum_{r=1}^{R} \log p(X^r, Q^r) \\
&= \sum_{r=1}^{R} \log \left[p(q_1^r) p(x_1^r | q_1^r) \prod_{t=2}^{T_r} p(q_t^r | q_{t-1}^r) p(x_t^r | q_t^r) \right] \\
&= \sum_{r=1}^{R} \left[\log \pi_{q_1^r} + \log b_{q_1^r}(x_1^r) + \sum_{t=2}^{T_r} \left(\log a_{q_t^r q_{t-1}^r} + \log b_{q_t^r}(x_t^r) \right) \right]
\end{aligned}
$$

Optimization with Constraints

$$\begin{aligned}
\max_{\pi} \quad & \sum_{r=1}^{R} \log p(X^r, Q^r) \\
\text{s.t.} \quad & \sum_{j=1}^{J} \pi_j = 1 \\
& 0 \leq \pi_j \leq 1 \quad \text{for } j = 1, \ldots, J
\end{aligned}$$

Lagrangian

$$\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta \left(\sum_{j=1}^{J} \pi_j - 1 \right)$$

Lagrangian

$$\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta \left(\sum_{j=1}^{J} \pi_j - 1 \right)$$

- Suppose $\sum_{j=1}^{J} \pi_i \geq 1$.

Lagrangian

$$\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta \left(\sum_{j=1}^{J} \pi_j - 1 \right)$$

- Suppose $\sum_{j=1}^{J} \pi_i \geq 1$.
- If $\eta = -\infty$, then $\sum_{j=1}^{J} \pi_i$ has to be 1 for the objective not to be $-\infty$.

Lagrangian

$$\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta \left(\sum_{j=1}^{J} \pi_j - 1 \right)$$

- Suppose $\sum_{j=1}^{J} \pi_i \geq 1$.
- If $\eta = -\infty$, then $\sum_{j=1}^{J} \pi_i$ has to be 1 for the objective not to be $-\infty$.
- In general, $\eta < 0$, and we get penalized if the constraint is not satisfied.

$$\frac{\partial}{\partial \pi_i} \left[\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta \left(\sum_{j=1}^{J} \pi_j - 1 \right) \right]$$

$$\frac{\partial}{\partial \pi_i}\left[\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta\left(\sum_{j=1}^{J} \pi_j - 1\right)\right] = \sum_{r=1}^{R} \frac{\mathbb{1}_{q_1^r = i}}{\pi_i} + \eta = 0$$

$$\frac{\partial}{\partial \pi_i} \left[\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta \left(\sum_{j=1}^{J} \pi_j - 1 \right) \right] = \sum_{r=1}^{R} \frac{\mathbb{1}_{q_1^r = i}}{\pi_i} + \eta = 0$$

$$\implies \pi_i = -\frac{1}{\eta} \sum_{r=1}^{R} \mathbb{1}_{q_1^r = i}$$

$$\frac{\partial}{\partial \pi_i}\left[\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta\left(\sum_{j=1}^{J} \pi_j - 1\right)\right] = \sum_{r=1}^{R} \frac{\mathbb{1}_{q_1^r = i}}{\pi_i} + \eta = 0$$

$$\implies \pi_i = -\frac{1}{\eta}\sum_{r=1}^{R} \mathbb{1}_{q_1^r = i}$$

$$\sum_{j=1}^{J} \pi_j = \sum_{j=1}^{J} -\frac{1}{\eta}\sum_{r=1}^{R} \mathbb{1}_{q_1^r = j} = 1$$

$$\frac{\partial}{\partial \pi_i}\left[\sum_{r=1}^{R}\log p(X^r, Q^r) + \eta\left(\sum_{j=1}^{J}\pi_j - 1\right)\right] = \sum_{r=1}^{R}\frac{\mathbb{1}_{q_1^r=i}}{\pi_i} + \eta = 0$$

$$\implies \pi_i = -\frac{1}{\eta}\sum_{r=1}^{R}\mathbb{1}_{q_1^r=i}$$

$$\sum_{j=1}^{J}\pi_j = \sum_{j=1}^{J} -\frac{1}{\eta}\sum_{r=1}^{R}\mathbb{1}_{q_1^r=j} = 1 \implies \eta = -\sum_{r=1}^{R}\sum_{j=1}^{J}\mathbb{1}_{q_1^r=j}$$

$$\frac{\partial}{\partial \pi_i}\left[\sum_{r=1}^{R} \log p(X^r, Q^r) + \eta\left(\sum_{j=1}^{J} \pi_j - 1\right)\right] = \sum_{r=1}^{R} \frac{\mathbb{1}_{q_1^r = i}}{\pi_i} + \eta = 0$$

$$\implies \pi_i = -\frac{1}{\eta}\sum_{r=1}^{R} \mathbb{1}_{q_1^r = i} = \frac{\sum_{r=1}^{R} \mathbb{1}_{q_1^r = i}}{\sum_{r=1}^{R}\sum_{j=1}^{J} \mathbb{1}_{q_1^r = j}}$$

$$\sum_{j=1}^{J} \pi_j = \sum_{j=1}^{J} -\frac{1}{\eta}\sum_{r=1}^{R} \mathbb{1}_{q_1^r = j} = 1 \implies \eta = -\sum_{r=1}^{R}\sum_{j=1}^{J} \mathbb{1}_{q_1^r = j}$$

Viterbi Training

- What would be the best parameters had we know the hidden sequence?

$$\lambda^{s+1} = \operatorname*{argmax}_{\lambda} \prod_{r=1}^{R} p_\lambda(X^r, \hat{Q}^r)$$

Viterbi Training

- What would be the best parameters had we know the hidden sequence?

$$\lambda^{s+1} = \operatorname*{argmax}_{\lambda} \prod_{r=1}^{R} p_{\lambda}(X^r, \hat{Q}^r)$$

- What would be the best hidden sequence had we know the parameters?

$$\hat{Q}^r = \operatorname*{argmax}_{Q} p_{\lambda^s}(X^r, Q)$$

Viterbi Training

- $\lambda^1, \lambda^2, \ldots, \lambda^s$

Viterbi Training

- $\lambda^1, \lambda^2, \ldots, \lambda^s$
- Does Viterbi training converge?
- Does Viterbi training improve $\prod_{r=1}^{R} p(X^r)$?

Viterbi Training

- $\lambda^1, \lambda^2, \ldots, \lambda^s$
- Does Viterbi training converge?
- Does Viterbi training improve $\prod_{r=1}^{R} p(X^r)$?
- Instead of using one hidden sequence, could we use all of them?

- Training with one hidden sequence

$$\lambda^{s+1} = \underset{\lambda}{\text{argmax}} \sum_{r=1}^{R} \log p_{\lambda}(X^r, \hat{Q}^r)$$

$$\text{where } \hat{Q}^r = \underset{Q}{\text{argmax}}\, p_{\lambda^s}(X^r, Q)$$

- Training with one hidden sequence

$$\lambda^{s+1} = \operatorname*{argmax}_{\lambda} \sum_{r=1}^{R} \log p_{\lambda}(X^r, \hat{Q}^r)$$

$$\text{where } \hat{Q}^r = \operatorname*{argmax}_{Q} p_{\lambda^s}(X^r, Q)$$

- Training with all hidden sequences

$$\lambda^{s+1} = \operatorname*{argmax}_{\lambda} \sum_{r=1}^{R} \mathbb{E}_{Q \sim p_{\lambda^s}(Q|X^r)}[\log p_{\lambda}(X^r, Q)]$$

Expectation Maximization

- E-step: Compute $\mathbb{E}_{Q \sim p_{\lambda^s}(Q|X^r)}[\log p_{\lambda}(X^r, Q)]$ for $r = 1, \ldots, R$.
- M-step:

$$\lambda^{s+1} = \operatorname*{argmax}_{\lambda} \sum_{r=1}^{R} \mathbb{E}_{Q \sim p_{\lambda^s}(Q|X^r)}[\log p_{\lambda}(X^r, Q)]$$

Computing the EM Objective

$$\mathbb{E}_{Q \sim p(Q|X)}[\log p(X, Q)] = \sum_{Q} p(Q|X)[\log p(X, Q)]$$

Computing the EM Objective

$$\mathbb{E}_{Q \sim p(Q|X)}[\log p(X, Q)] = \sum_{Q} p(Q|X)[\log p(X, Q)]$$

$$= \frac{1}{p(x_{1:T})} \sum_{q_{1:T}} p(x_{1:T}, q_{1:T}) \log p(x_{1:T}, q_{1:T})$$

$$\alpha'_{q_t}(t) = \sum_{q_{1:t-1}} p(x_{1:t}, q_{1:t}) \log p(x_{1:t}, q_{1:t})$$

$$
\begin{aligned}
\alpha'_{q_t}(t) &= \sum_{q_{1:t-1}} p(x_{1:t}, q_{1:t}) \log p(x_{1:t}, q_{1:t}) \\
&= \sum_{q_{t-1}} \sum_{q_{1:t-2}} p(x_{1:t-1}, q_{1:t-1}) p(q_t | q_{t-1}) p(x_t | q_t) \\
&\qquad \left[\log p(x_{1:t-1}, q_{1:t-1}) + \log p(q_t | q_{t-1}) p(x_t | q_t) \right]
\end{aligned}
$$

$$\alpha'_{q_t}(t) = \sum_{q_{1:t-1}} p(x_{1:t}, q_{1:t}) \log p(x_{1:t}, q_{1:t})$$

$$= \sum_{q_{t-1}} \sum_{q_{1:t-2}} p(x_{1:t-1}, q_{1:t-1}) p(q_t | q_{t-1}) p(x_t | q_t) \left[\log p(x_{1:t-1}, q_{1:t-1}) + \log p(q_t | q_{t-1}) p(x_t | q_t) \right]$$

$$
\begin{aligned}
\alpha'_{q_t}(t) &= \sum_{q_{1:t-1}} p(x_{1:t}, q_{1:t}) \log p(x_{1:t}, q_{1:t}) \\
&= \sum_{q_{t-1}} \sum_{q_{1:t-2}} p(x_{1:t-1}, q_{1:t-1}) p(q_t|q_{t-1}) p(x_t|q_t) \\
&\qquad \left[\log p(x_{1:t-1}, q_{1:t-1}) + \log p(q_t|q_{t-1}) p(x_t|q_t) \right] \\
&= \sum_{q_{t-1}} p(q_t|q_{t-1}) p(x_t|q_t) \\
&\qquad \left[\alpha'_{q_{t-1}}(t-1) + \alpha_{q_{t-1}}(t-1) \log p(q_t|q_{t-1}) p(x_t|q_t) \right]
\end{aligned}
$$

$$
\begin{aligned}
\alpha'_{q_t}(t) &= \sum_{q_{1:t-1}} p(x_{1:t}, q_{1:t}) \log p(x_{1:t}, q_{1:t}) \\
&= \sum_{q_{t-1}} \sum_{q_{1:t-2}} p(x_{1:t-1}, q_{1:t-1}) p(q_t|q_{t-1}) p(x_t|q_t) \\
&\qquad \left[\log p(x_{1:t-1}, q_{1:t-1}) + \log p(q_t|q_{t-1}) p(x_t|q_t) \right] \\
&= \sum_{q_{t-1}} p(q_t|q_{t-1}) p(x_t|q_t) \\
&\qquad \left[\alpha'_{q_{t-1}}(t-1) + \alpha_{q_{t-1}}(t-1) \log p(q_t|q_{t-1}) p(x_t|q_t) \right]
\end{aligned}
$$

$$
\begin{aligned}
\alpha'_{q_t}(t) &= \sum_{q_{1:t-1}} p(x_{1:t}, q_{1:t}) \log p(x_{1:t}, q_{1:t}) \\
&= \sum_{q_{t-1}} \sum_{q_{1:t-2}} p(x_{1:t-1}, q_{1:t-1}) p(q_t|q_{t-1}) p(x_t|q_t) \\
&\qquad \left[\log p(x_{1:t-1}, q_{1:t-1}) + \log p(q_t|q_{t-1}) p(x_t|q_t) \right] \\
&= \sum_{q_{t-1}} p(q_t|q_{t-1}) p(x_t|q_t) \\
&\qquad \left[\alpha'_{q_{t-1}}(t-1) + \alpha_{q_{t-1}}(t-1) \log p(q_t|q_{t-1}) p(x_t|q_t) \right]
\end{aligned}
$$

$$
\mathbb{E}_{Q \sim p(Q|X)}[\log p(X, Q)] = \sum_{Q} p(Q|X)[\log p(X, Q)] = \sum_{q_T} \alpha'_{q_T}(T)
$$

M-Step

$$\frac{\partial}{\partial \pi_i} \sum_{r=1}^{R} \sum_{Q^r} p(Q^r | X^r) \log p_\pi(X^r, Q^r)$$

M-Step

$$\frac{\partial}{\partial \pi_i} \sum_{r=1}^{R} \sum_{Q^r} p(Q^r|X^r) \log p_\pi(X^r, Q^r)$$

$$= \sum_{r=1}^{R} \sum_{Q^r} p(Q^r|X^r) \frac{\partial}{\partial \pi_i} \log p_\pi(X^r, Q^r)$$

M-Step

$$\frac{\partial}{\partial \pi_i} \sum_{r=1}^{R} \sum_{Q^r} p(Q^r|X^r) \log p_\pi(X^r, Q^r)$$

$$= \sum_{r=1}^{R} \sum_{Q^r} p(Q^r|X^r) \frac{\partial}{\partial \pi_i} \log p_\pi(X^r, Q^r)$$

M-Step

$$
\begin{aligned}
&\frac{\partial}{\partial \pi_i} \sum_{r=1}^{R} \sum_{Q^r} p(Q^r | X^r) \log p_\pi(X^r, Q^r) \\
&= \sum_{r=1}^{R} \sum_{Q^r} p(Q^r | X^r) \frac{\partial}{\partial \pi_i} \log p_\pi(X^r, Q^r) \\
&= \sum_{r=1}^{R} \sum_{q_1^r} p(q_1^r | X^r) \frac{\mathbb{1}_{q_1^r = i}}{\pi_i}
\end{aligned}
$$

M-Step

$$\frac{\partial}{\partial \pi_i} \sum_{r=1}^{R} \sum_{Q^r} p(Q^r | X^r) \log p_\pi(X^r, Q^r)$$

$$= \sum_{r=1}^{R} \sum_{Q^r} p(Q^r | X^r) \frac{\partial}{\partial \pi_i} \log p_\pi(X^r, Q^r)$$

$$= \sum_{r=1}^{R} \sum_{q_1^r} p(q_1^r | X^r) \frac{\mathbb{1}_{q_1^r = i}}{\pi_i}$$

M-Step

$$
\begin{aligned}
&\frac{\partial}{\partial \pi_i} \sum_{r=1}^{R} \sum_{Q^r} p(Q^r | X^r) \log p_\pi(X^r, Q^r) \\
&= \sum_{r=1}^{R} \sum_{Q^r} p(Q^r | X^r) \frac{\partial}{\partial \pi_i} \log p_\pi(X^r, Q^r) \\
&= \sum_{r=1}^{R} \sum_{q_1^r} p(q_1^r | X^r) \frac{\mathbb{1}_{q_1^r = i}}{\pi_i} \\
&= \sum_{r=1}^{R} p(q_1^r = i | X^r) \frac{1}{\pi_i}
\end{aligned}
$$

M-Step

$$\pi_i^{s+1} = \frac{\sum_{r=1}^{R} p(q_1^r = i|X^r)}{\sum_{r=1}^{R} \sum_{j=1}^{J} p(q_1^r = j|X^r)}$$

M-Step

$$\pi_i^{s+1} = \frac{\sum_{r=1}^{R} p(q_1^r = i | X^r)}{\sum_{r=1}^{R} \sum_{j=1}^{J} p(q_1^r = j | X^r)}$$

$$p(q_1^r = i | X^r) = \frac{p(X^r | q_1^r = i) p(q_1^r = i)}{p(X^r)} = \frac{\beta_i^r(1) \pi_i^s}{p(X^r)}$$

M-Step

$$\pi_i^{s+1} = \frac{\sum_{r=1}^{R} p(q_1^r = i|X^r)}{\sum_{r=1}^{R} \sum_{j=1}^{J} p(q_1^r = j|X^r)} = \frac{\sum_{r=1}^{R} \beta_i^r(1)\pi_i^s}{\sum_{r=1}^{R} \sum_{j=1}^{J} \beta_j^r(1)\pi_j^s}$$

$$p(q_1^r = i|X^r) = \frac{p(X^r|q_1^r = i)p(q_1^r = i)}{p(X^r)} = \frac{\beta_i^r(1)\pi_i^s}{p(X^r)}$$

- Viterbi Training

$$\pi_i^{s+1} = \frac{\sum_{r=1}^{R} \mathbb{1}_{q_1^r = i}}{\sum_{r=1}^{R} \sum_{j=1}^{J} \mathbb{1}_{q_1^r = j}}$$

- EM

$$\pi_i^{s+1} = \frac{\sum_{r=1}^{R} \beta_i^r(1)\pi_i^s}{\sum_{r=1}^{R} \sum_{j=1}^{J} \beta_j^r(1)\pi_j^s}$$

$$\gamma_j(t) = p(q_t = j|X) = \frac{\alpha_j(t)\beta_j(t)}{p(X)}$$

$$\gamma_j(t) = p(q_t = j|X) = \frac{\alpha_j(t)\beta_j(t)}{p(X)}$$

$$\xi_{i,j}(t) = p(q_t = j, q_{t-1} = i|X) = \frac{\alpha_i(t-1)a_{ij}^s b_j(x_t)\beta_j(t)}{p(X)}$$

$$a_{ij}^{s+1} = \frac{\sum_{r=1}^{R}\sum_{t=2}^{T_r}\xi_{i,j}^r(t)}{\sum_{r=1}^{R}\sum_{t=2}^{T_r}\sum_{k=1}^{J}\xi_{i,k}^r(t)}$$

Expectation Maximization

- E-step: Compute $\mathbb{E}_{Q \sim p_{\lambda^s}(Q|X^r)}[\log p_\lambda(X^r, Q)]$ for $r = 1, \dots, R$.

- M-step:

$$\lambda^{s+1} = \operatorname*{argmax}_{\lambda} \sum_{r=1}^{R} \mathbb{E}_{Q \sim p_{\lambda^s}(Q|X^r)}[\log p_\lambda(X^r, Q)]$$

Expectation Maximization

- E-step: Compute $\alpha, \beta, \alpha', \gamma, \xi$.
- M-step: Compute λ in closed form.

Expectation Maximization

- $\lambda^1, \lambda^2, \ldots, \lambda^s$
- Does EM converge?
- Does EM improve $\prod_{r=1}^{R} p(X^r)$?

Summary

- Stochastic Gradient Descent
- Viterbi Training
- Expectation Maximization

Further Reading

- Chapter 6, Rabiner and Juang, "Fundamentals of Speech Recognition," 1993
- Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," 1998
- Chapter 3, Beal, "Variational Algorithms for Approximate Bayesian Inference," 2003