Machine Learning

4771

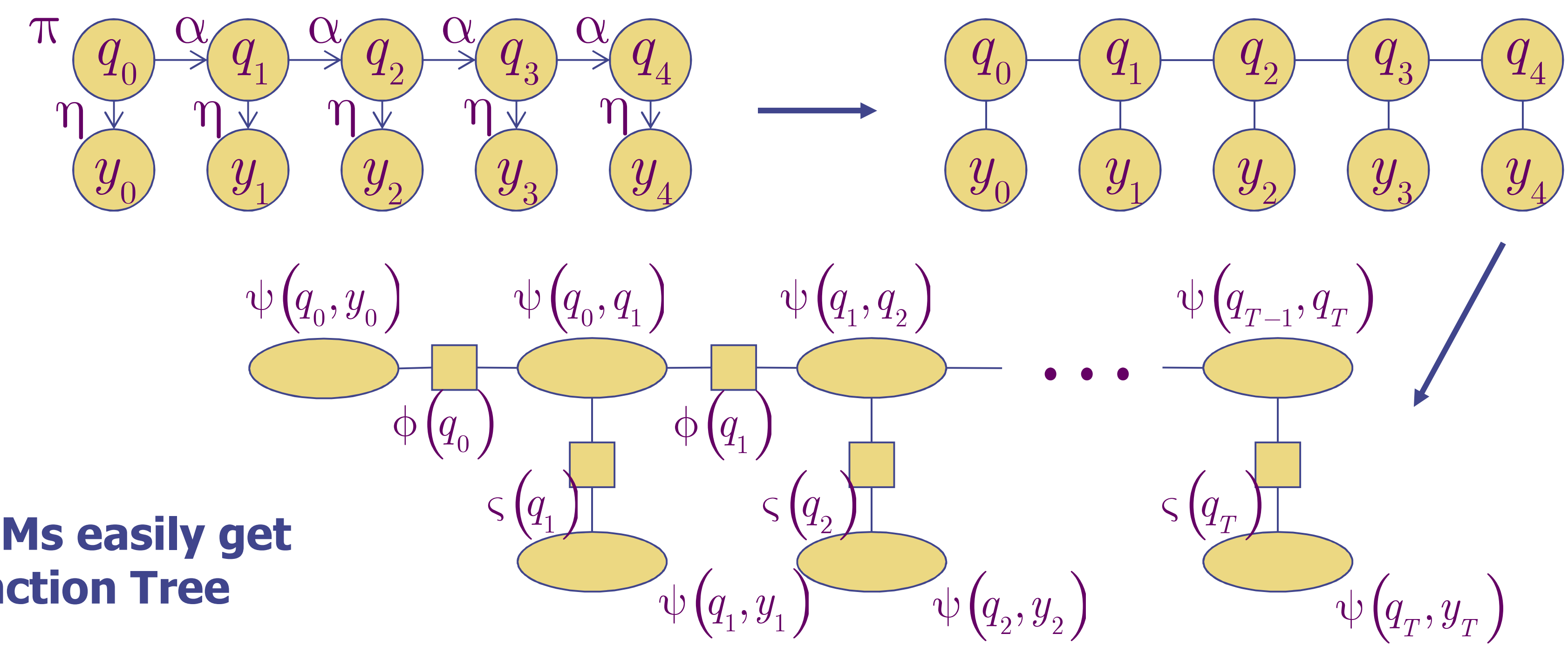
Instructors:
Adrian Weller and Ilia Vovsha

Lecture 25

- HMMs with Evidence
- HMM Collect
- HMM Evaluate
- HMM Distribute
- HMM Decode
- HMM Parameter Learning via JTA & EM

Recall HMM Basic Operations

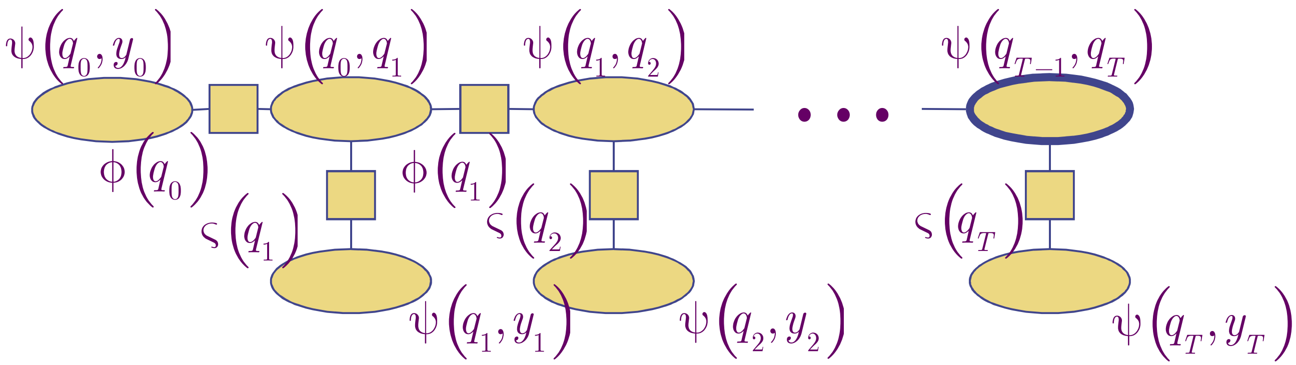
- Would like to do 3 basic things with our HMMs:
  1) Evaluate: given $y_0,\ldots,y_T$ & θ compute $p(y_1,\ldots,y_T)$
  2) Decode/inference: given $y_0,\ldots,y_T$ & θ find MAP $q_0,\ldots,q_T$ or marginals $p(q_0),\ldots,p(q_T)$
  3) Max Likelihood Learn: given $y_0,\ldots,y_T$ learn parameters θ
- Typically use Baum-Welch (α–β algo)… JTA is more general:



HMMs easily get Junction Tree

HMMs: JTA Init & Verify

•Init: $\psi(q_0,y_0) = p(q_0)p(y_0 \mid q_0) \quad \psi(q_t,q_{t+1}) = p(q_{t+1} \mid q_t) = \alpha_{q_t,q_{t+1}} \quad \psi(q_t,y_t) = p(y_t \mid q_t)$



$Z = 1$

$\phi(q_t) = 1$

$\zeta(q_t) = 1$

•Collect *up* from leaves: doesn't change zeta separators

$$\zeta^*(q_t) = \sum_{y_t} \psi(q_t,y_t) = \sum_{y_t} p(y_t \mid q_t) = 1 \qquad \psi^*(q_{t-1},q_t) = \frac{\zeta^*}{\zeta}\psi(q_{t-1},q_t) = \psi(q_{t-1},q_t)$$

•Collect *left-right* via phi's: changes backbone to marginals

$$\phi^*(q_0) = \sum_{y_0} \psi(q_0,y_0) = p(q_0) \qquad \psi^*(q_0,q_1) = \frac{\phi^*}{\phi}\psi(q_0,q_1) = p(q_0,q_1)$$

$$\phi^*(q_t) = \sum_{q_{t-1}} \psi^*(q_{t-1},q_t) = p(q_t) \qquad \psi^*(q_{t-1},q_t) = \frac{p(q_{t-1})}{1} p(q_t \mid q_{t-1}) = p(q_{t-1},q_t)$$

•Distribute:

$$\zeta^{**}(q_t) = \sum_{q_{t-1}} \psi^*(q_{t-1},q_t) = \sum_{q_{t-1}} p(q_{t-1},q_t) = p(q_t)$$

$$\psi^{**}(q_t,y_t) = \frac{\zeta^{**}}{\zeta^*}\psi(q_t,y_t) = \frac{p(q_t)}{1} p(y_t \mid q_t) = p(y_t,q_t)$$

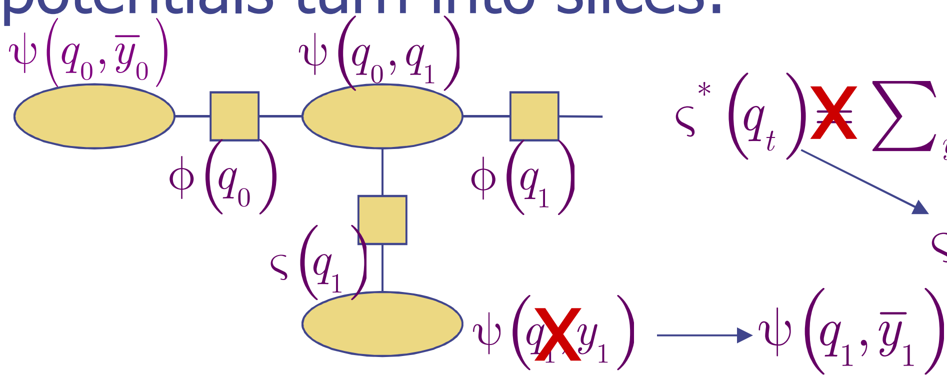
...done!

HMMs: JTA with Evidence

- If y sequence is observed (in problems 1,2,3) get evidence:

$$p(q,\bar{y}) = p(q_0)\prod_{t=1}^{T} p(q_t \mid q_{t-1})\prod_{t=0}^{T} p(\bar{y}_t \mid q_t)$$

- The potentials turn into slices:



$$\varsigma^*(q_t) \cancel{=} \sum_{y_t} \psi(q_t,y_t)$$

$$\varsigma^*(q_t) = \psi(q_t,\bar{y}_t) = p(\bar{y}_t \mid q_t)$$

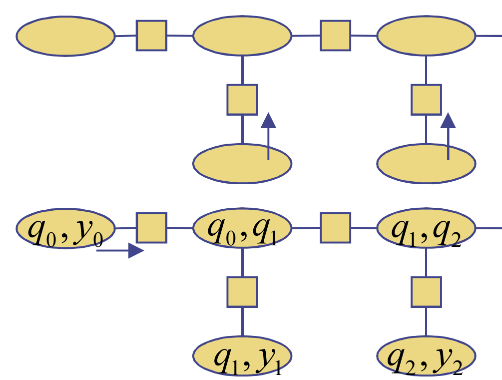
- Next, pick a root, for example *rightmost* one: $\psi(q_{T-1},q_T)$

- Collect all zeta separators bottom up:

$$\varsigma^*(q_t) = \psi(q_t,\bar{y}_t) = p(\bar{y}_t \mid q_t)$$

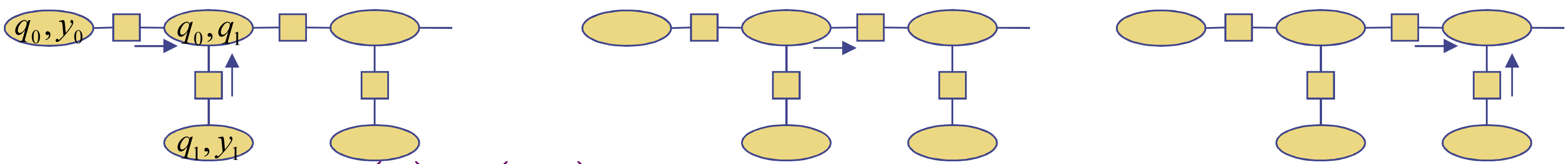
- Collect leftmost phi separator to the right:

$$\phi^*(q_0) = \sum_{y_0} \psi(q_0,\bar{y}_0)\delta(y_0-\bar{y}_0) = p(\bar{y}_0,q_0)$$



HMMs: Collect with Evidence

- Now, we will collect (*) along the backbone left to right
- Update each clique with its left and bottom separators:



$$\psi^*\left(q_t, q_{t+1}\right) = \frac{\phi^*\left(q_t\right)}{1} \frac{\varsigma^*\left(q_{t+1}\right)}{1} \psi\left(q_t, q_{t+1}\right) = \phi^*\left(q_t\right) p\left(\bar{y}_{t+1} \mid q_{t+1}\right) \alpha_{q_t, q_{t+1}}$$

$$\phi^*\left(q_{t+1}\right) = \sum_{q_t} \psi^*\left(q_t, q_{t+1}\right) = \sum_{q_t} \phi^*\left(q_t\right) p\left(\bar{y}_{t+1} \mid q_{t+1}\right) \alpha_{q_t, q_{t+1}}$$

- Keep going along chain until right most node
- Note: above formula for phi is recursive, could use as is.
- Recall we had $\phi^*\left(q_0\right) = p\left(\bar{y}_0, q_0\right)$
- Hence

$$\begin{aligned} \psi^*(q_0, q_1) &= p\left(\bar{y}_0, q_0\right) p\left(\bar{y}_1 \mid q_1\right) p\left(q_1 \mid q_0\right) \\ &= p\left(\bar{y}_0, q_0\right) p\left(\bar{y}_1 \mid q_1, q_0, \bar{y}_0\right) p\left(q_1 \mid q_0, \bar{y}_0\right) \quad \text{conditional indep} \\ &= p\left(\bar{y}_0, q_0\right) p\left(\bar{y}_1, q_1 \mid q_0, \bar{y}_0\right) = p\left(\bar{y}_0, \bar{y}_1, q_0, q_1\right) \end{aligned}$$

HMMs: Evaluate with Evidence

Hence $\phi^*(q_1) = \sum_{q_0} \psi^*(q_0, q_1) = \sum_{q_0} p(\bar{y}_0, \bar{y}_1, q_0, q_1) = p(\bar{y}_0, \bar{y}_1, q_1)$

Continuing, we obtain…

$$\psi^*(q_{t-1}, q_t) = p(\bar{y}_0, \ldots, \bar{y}_t, q_{t-1}, q_t)$$

$$\phi^*(q_t) = \sum_{q_{t-1}} \psi^*(q_{t-1}, q_t) = \sum_{q_{t-1}} p(\bar{y}_0, \ldots, \bar{y}_t, q_{t-1}, q_t) = p(\bar{y}_0, \ldots, \bar{y}_t, q_t)$$

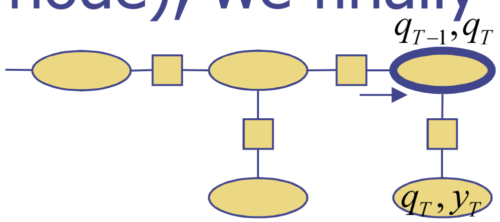
These are the marginals AND observed data up to the point t

HMMs: Evaluate with Evidence

- If we are solving the first HMM problem, likelihood:
  1) Evaluate: given $y_0,\ldots,y_T$ & θ, compute $p(y_0,\ldots,y_T|\theta)$
- We are already almost done! Collect is enough.
- As we collect to the root (rightmost node), we finally get:

$$\psi^*(q_{T-1},q_T) = p(\bar{y}_0,\ldots,\bar{y}_T,q_{T-1},q_T)$$

marginal AND all observed data



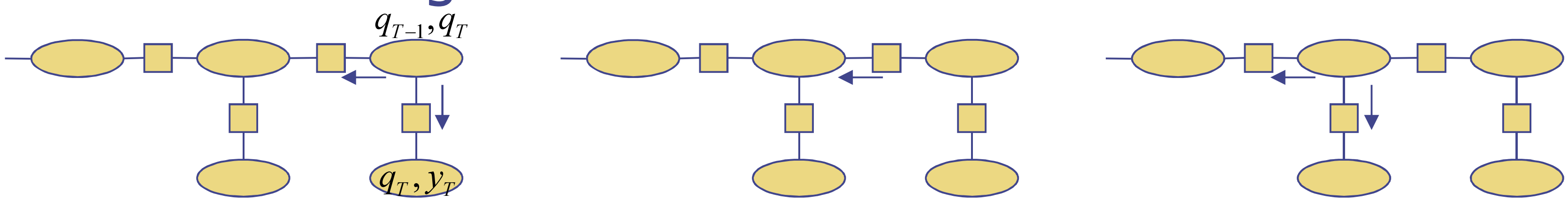
- Can compute the likelihood just by summing this root ψ^*

$$p\left(\bar{y}_0,\ldots,\bar{y}_T\right) = \sum_{q_{T-1},q_T} p\left(\bar{y}_0,\ldots,\bar{y}_T,q_{T-1},q_T\right) = \sum_{q_{T-1},q_T} \psi^*(q_{T-1},q_T)$$

- What should we expect in the distribute phase?

HMMs: Distribute with Evidence

- Now, we distribute (**) along the backbone right to left
- Have first ** for root (stays the same): $\psi^{**}\left(q_{T-1},q_T\right)=\psi^{*}\left(q_{T-1},q_T\right)$
- Start distributing from there:



$$\phi^{**}\left(q_t\right)=\sum_{q_{t+1}}\psi^{**}\left(q_t,q_{t+1}\right)$$

$$\psi^{**}\left(q_t,q_{t+1}\right)=\frac{\phi^{**}\left(q_{t+1}\right)}{\phi^{*}\left(q_{t+1}\right)}\psi^{*}\left(q_t,q_{t+1}\right)$$

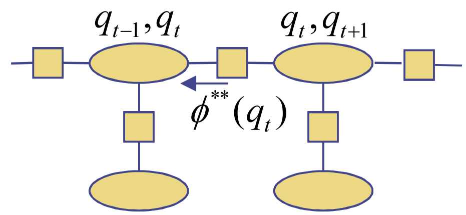
$$\varsigma^{**}\left(q_t\right)=\sum_{q_{t-1}}\psi^{**}\left(q_{t-1},q_t\right)$$

$$\begin{aligned}\psi^{**}(q_T,y_T)&=\frac{\varsigma^{**}\left(q_T\right)}{p\left(\bar{y}_T\mid q_T\right)}p\left(\bar{y}_T\mid q_T\right)\\&=p(\bar{y}_0,\ldots,\bar{y}_T,q_T)\end{aligned}$$

$$\begin{aligned}\psi^{**}(q_{T-2},q_{T-1})&=\frac{\varphi^{**}\left(q_{T-1}\right)}{\varphi^{*}\left(q_{T-1}\right)}p(\bar{y}_0,\ldots,\bar{y}_{T-1},q_{T-2},q_{T-1})\\&=\frac{p(\bar{y}_0,\ldots,\bar{y}_T,q_{T-1})}{p(\bar{y}_0,\ldots,\bar{y}_{T-1},q_{T-1})}p(\bar{y}_0,\ldots,\bar{y}_{T-1},q_{T-2},q_{T-1})\overset{\text{why?}}{=}p(\bar{y}_0,\ldots,\bar{y}_{T-1},\bar{y}_T,q_{T-2},q_{T-1})\end{aligned}$$

HMMs: Distribute with Evidence

Examine the general case:



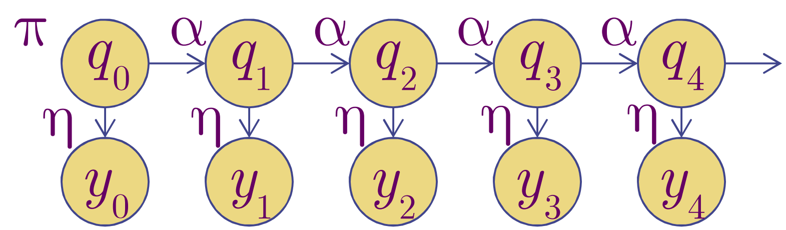
Assume $\psi^{**}(q_t, q_{t+1}) = p(\bar{y}_0, \ldots, \bar{y}_T, q_t, q_{t+1})$ [true for $t+1 = T$]

Then $\phi^{**}(q_t) = \sum_{q_{t+1}} \psi^{**}(q_t, q_{t+1}) = p(\bar{y}_0, \ldots, \bar{y}_T, q_t)$

$$
\begin{aligned}
\psi^{**}(q_{t-1}, q_t) &= \frac{\varphi^{**}(q_t)}{\varphi^{*}(q_t)} \psi^{*}(q_{t-1}, q_t) = \frac{p(\bar{y}_0, \ldots, \bar{y}_T, q_t)}{p(\bar{y}_0, \ldots, \bar{y}_t, q_t)} p(\bar{y}_0, \ldots, \bar{y}_t, q_{t-1}, q_t) \\
&= p(\bar{y}_{t+1}, \ldots, \bar{y}_T \mid \bar{y}_0, \ldots, \bar{y}_t, q_t) p(\bar{y}_0, \ldots, \bar{y}_t, q_{t-1}, q_t) \\
&= p(\bar{y}_{t+1}, \ldots, \bar{y}_T \mid \bar{y}_0, \ldots, \bar{y}_t, q_{t-1}, q_t) p(\bar{y}_0, \ldots, \bar{y}_t, q_{t-1}, q_t) \\
&= p(\bar{y}_0, \ldots, \bar{y}_t, \ldots, \bar{y}_T, q_{t-1}, q_t)
\end{aligned}
$$

conditional independence, see below

Hence result holds for all t



Original directed model

HMMs: Marginals & MaxDecoding

- After JTA is finished, we have the following:

$$\phi^{**}\left(q_t\right) \propto p\left(q_t \mid \bar{y}_1,\ldots,\bar{y}_T\right) \qquad \varsigma^{**}\left(q_{t+1}\right) \propto p\left(q_{t+1} \mid \bar{y}_1,\ldots,\bar{y}_T\right)$$

$$\psi^{**}\left(q_t, q_{t+1}\right) \propto p\left(q_t, q_{t+1} \mid \bar{y}_1,\ldots,\bar{y}_T\right)$$

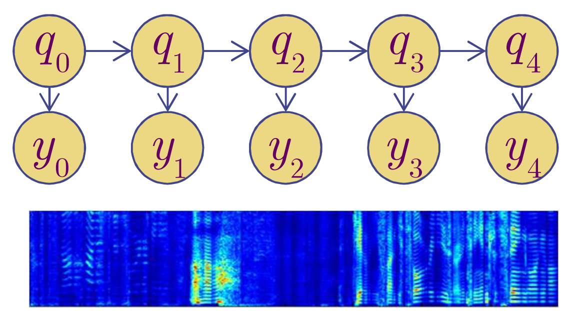
(normalize to get these conditionals)

- We have solved part of the HMM Problem:

 2) Decode: given $y_0,\ldots,y_T$ & θ find $p(q_0),\ldots,p(q_T)$ and $q_0,\ldots,q_T$

- The separators define a distribution over the hidden states
- e.g. the probability the audio y_t was due to phoneme q_t
- We can also decode to find the most likely path $q_0 \ldots q_T$
- Here, we use the ArgMax JTA algorithm
- Run JTA but replace sums with max
- Then, find biggest entry in separators:

$$\hat{q}_t = \arg\max_{q_t} \phi^{**}\left(q_t\right) \quad \forall t = 0\ldots T$$



HMMs: EM Learning

- Finally 3) Max Likelihood: given $y_0,\ldots,y_T$ learn parameters θ
- Recall max likelihood: $\hat{\theta} = \arg\max_\theta \log p\left(\bar{y} \mid \theta\right)$
- If observe q, it's easy to maximize the *complete* likelihood:

$$
\begin{aligned}
l\left(\theta\right) &= \log\left(p\left(q,y\right)\right)\\
&= \log\left(p\left(q_0\right)\prod_{t=1}^{T} p\left(q_t \mid q_{t-1}\right)\prod_{t=0}^{T} p\left(\bar{y}_t \mid q_t\right)\right)\\
&= \log p\left(q_0\right) + \sum_{t=1}^{T}\log p\left(q_t \mid q_{t-1}\right) + \sum_{t=0}^{T}\log p\left(\bar{y}_t \mid q_t\right)\\
&= \log\prod_{i=1}^{M}\left[\pi_i\right]^{q_0^i} + \sum_{t=1}^{T}\log\prod_{i=1}^{M}\prod_{j=1}^{M}\left[\alpha_{ij}\right]^{q_{t-1}^i q_t^j} + \sum_{t=0}^{T}\log\prod_{i=1}^{M}\prod_{j=1}^{N}\left[\eta_{ij}\right]^{q_t^i y_t^j}\\
&= \sum_{i=1}^{M} q_0^i \log\pi_i + \sum_{t=1}^{T}\sum_{i,j=1}^{M} q_{t-1}^i q_t^j \log\alpha_{ij} + \sum_{t=0}^{T}\sum_{i=1}^{M}\sum_{j=1}^{N} q_t^i y_t^j \log\eta_{ij}
\end{aligned}
$$

Indicators, values in {0,1}

Introduce Lagrangian & take derivatives

$$\sum_{i=1}^{M}\pi_i = 1 \qquad \sum_{j=1}^{M}\alpha_{ij} = 1 \qquad \sum_{j=1}^{N}\eta_{ij} = 1$$

$$\hat{\pi}_i = q_0^i \qquad \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} q_t^i q_{t+1}^j}{\sum_{k=1}^{M}\sum_{t=0}^{T-1} q_t^i q_{t+1}^k} \qquad \hat{\eta}_{ij} = \frac{\sum_{t=0}^{T} q_t^i y_t^j}{\sum_{k=1}^{N}\sum_{t=0}^{T} q_t^i y_t^k}$$

HMMs: EM Learning

- But, we don't observe the q's, incomplete...

$$p\left(\bar{y} \mid \theta\right) = \sum_{q} p\left(q, \bar{y} \mid \theta\right) = \sum_{q_0} \cdots \sum_{q_T} p\left(q_0\right) \prod_{t=1}^{T} p\left(q_t \mid q_{t-1}\right) \prod_{t=0}^{T} p\left(\bar{y}_t \mid q_t\right)$$

- EM: Max expected complete likelihood given current p(q)

$$E\left\{l\left(\theta\right)\right\} = E_{p\left(q_0,\ldots,q_T \mid y\right)}\left\{\log p\left(q, y\right)\right\} = \sum_{q_0} \cdots \sum_{q_T} p\left(q \mid y\right) \log p\left(q, y\right)$$

$$= E\left\{\sum_{i=1}^{M} q_0^i \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} q_{t-1}^i q_t^j \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} q_t^i y_t^j \log \eta_{ij}\right\}$$

$$= \sum_{i=1}^{M} E\left\{q_0^i\right\} \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} E\left\{q_{t-1}^i q_t^j\right\} \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} E\left\{q_t^i\right\} y_t^j \log \eta_{ij}$$

- M-step is maximizing as before:

$$\hat{\pi}_i = E\left\{q_0^i\right\} \qquad \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} E\left\{q_t^i q_{t+1}^j\right\}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} E\left\{q_t^i q_{t+1}^k\right\}} \qquad \hat{\eta}_{ij} = \frac{\sum_{t=0}^{T} E\left\{q_t^i\right\} y_t^j}{\sum_{k=1}^{N} \sum_{t=0}^{T} E\left\{q_t^i\right\} y_t^k}$$

- What are E{}'s?

HMMs: EM Learning

- But, we don't observe the q's, incomplete...

$$p\left(\bar{y} \mid \theta\right) = \sum_q p\left(q, \bar{y} \mid \theta\right) = \sum_{q_0} \cdots \sum_{q_T} p\left(q_0\right) \prod_{t=1}^{T} p\left(q_t \mid q_{t-1}\right) \prod_{t=0}^{T} p\left(\bar{y}_t \mid q_t\right)$$

- EM: Max expected complete likelihood given current p(q)

$$E\left\{l\left(\theta\right)\right\} = E_{p\left(q_0,\ldots,q_T \mid y\right)}\left\{\log p\left(q, y\right)\right\} = \sum_{q_0} \cdots \sum_{q_T} p\left(q \mid y\right) \log p\left(q, y\right)$$

$$= E\left\{\sum_{i=1}^{M} q_0^i \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} q_{t-1}^i q_t^j \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} q_t^i y_t^j \log \eta_{ij}\right\}$$

$$= \sum_{i=1}^{M} E\left\{q_0^i\right\} \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} E\left\{q_{t-1}^i q_t^j\right\} \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} E\left\{q_t^i\right\} y_t^j \log \eta_{ij}$$

- M-step is maximizing as before:

$$\hat{\pi}_i = E\left\{q_0^i\right\} \qquad \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} E\left\{q_t^i q_{t+1}^j\right\}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} E\left\{q_t^i q_{t+1}^k\right\}} \qquad \hat{\eta}_{ij} = \frac{\sum_{t=0}^{T} E\left\{q_t^i\right\} y_t^j}{\sum_{k=1}^{N} \sum_{t=0}^{T} E\left\{q_t^i\right\} y_t^k}$$

- What are E{}'s? $E_{p(x)}\left\{x^i\right\} = \sum_x p\left(x\right) x^i = \sum_x p\left(x\right) \delta\left(x = x^i\right) = p\left(x^i\right)$

HMMs: EM Learning

- But, we don't observe the q's, incomplete...

$$p\left(\bar{y} \mid \theta\right) = \sum_q p\left(q, \bar{y} \mid \theta\right) = \sum_{q_0} \cdots \sum_{q_T} p\left(q_0\right) \prod_{t=1}^{T} p\left(q_t \mid q_{t-1}\right) \prod_{t=0}^{T} p\left(\bar{y}_t \mid q_t\right)$$

- EM: Max expected complete likelihood given current p(q)

$$E\left\{l\left(\theta\right)\right\} = E_{p\left(q_0,\ldots,q_T \mid y\right)}\left\{\log p\left(q, y\right)\right\} = \sum_{q_0} \cdots \sum_{q_T} p\left(q \mid y\right) \log p\left(q, y\right)$$

$$= E\left\{\sum_{i=1}^{M} q_0^i \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} q_{t-1}^i q_t^j \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} q_t^i y_t^j \log \eta_{ij}\right\}$$

$$= \sum_{i=1}^{M} E\left\{q_0^i\right\} \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{M} E\left\{q_{t-1}^i q_t^j\right\} \log \alpha_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} E\left\{q_t^i\right\} y_t^j \log \eta_{ij}$$

- M-step is maximizing as before:

$$\hat{\pi}_i = E\left\{q_0^i\right\} \qquad \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} E\left\{q_t^i q_{t+1}^j\right\}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} E\left\{q_t^i q_{t+1}^k\right\}} \qquad \hat{\eta}_{ij} = \frac{\sum_{t=0}^{T} E\left\{q_t^i\right\} y_t^j}{\sum_{k=1}^{N} \sum_{t=0}^{T} E\left\{q_t^i\right\} y_t^k}$$

- What are E{}'s? $E_{p(x)}\left\{x^i\right\} = \sum_x p\left(x\right) x^i = \sum_x p\left(x\right) \delta\left(x = x^i\right) = p\left(x^i\right)$
- Our JTA ψ & ϕ marginals! (JTA is the E-Step for given θ)

$$E\left\{q_t^i q_{t+1}^j\right\} = p\left(q_t = i, q_{t+1} = j \mid \bar{y}\right) \qquad E\left\{q_t^i\right\} = p\left(q_t = i \mid \bar{y}\right)$$

HMMs: Gaussian Emissions

- Instead of table for emissions, have Gaussian:

$$p\left(\bar{y} \mid \theta\right) = \sum_q p\left(q, \bar{y} \mid \theta\right) = \sum_{q_0} \cdots \sum_{q_T} p\left(q_0\right) \prod_{t=1}^{T} p\left(q_t \mid q_{t-1}\right) \prod_{t=0}^{T} p\left(\bar{y}_t \mid q_t\right)$$

where $p\left(\bar{y}_t \mid q_t\right) = N\left(\bar{y}_t \mid \mu_{q_t}, I\right)$

- Clique initialization: $\psi\left(q_t, \bar{y}_t\right) = \psi\left(q_t\right) = N\left(\bar{y}_t \mid \mu_{q_t}, I\right)$

- M-step is maximizing as before:

$$\hat{\pi}_i = E\left\{q_0^i\right\} \qquad \hat{\alpha}_{ij} = \frac{\sum_{t=0}^{T-1} E\left\{q_t^i q_{t+1}^j\right\}}{\sum_{k=1}^{M} \sum_{t=0}^{T-1} E\left\{q_t^i q_{t+1}^k\right\}} \qquad \vec{\mu}_i = \frac{\sum_{t=0}^{T} E\left\{q_t^i\right\} \bar{y}_t}{\sum_{t=0}^{T} E\left\{q_t^i\right\}}$$

- Can thus handle continuous time series as in speech recognition