

# Machine Learning

# Topic 20

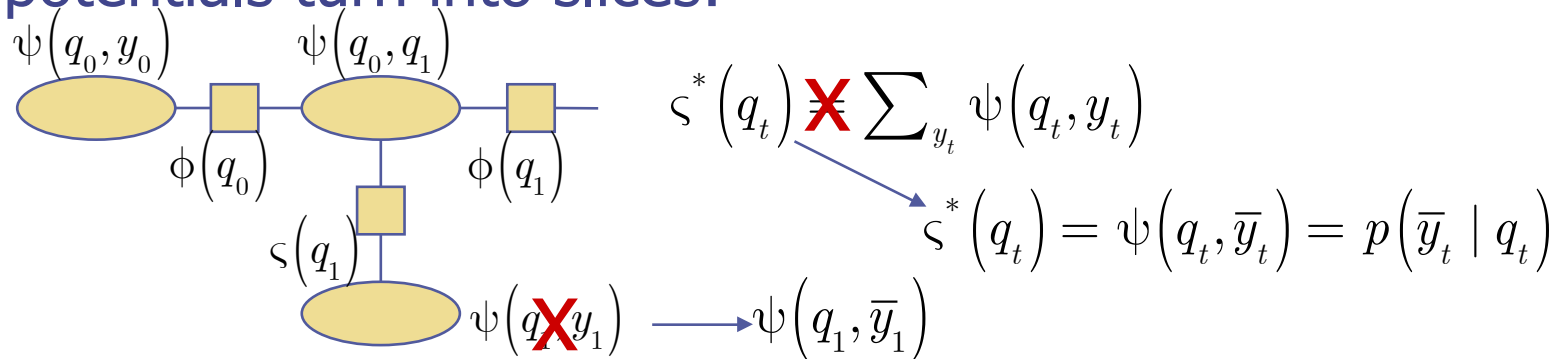
- HMMs with Evidence
- HMM Collect
- HMM Evaluate
- HMM Distribute
- HMM Decode
- HMM Parameter Learning via JTA & EM

# HMMs: JTA with Evidence

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

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

- The potentials turn into slices:



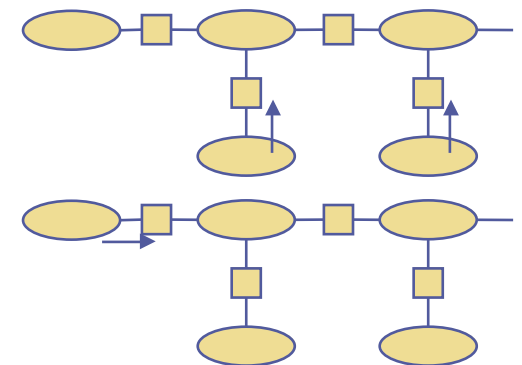
- Next, pick a root, for example *rightmost* one:  $\psi(q_{T-1}, q_T)$

- Collect all zeta separators bottom up:

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

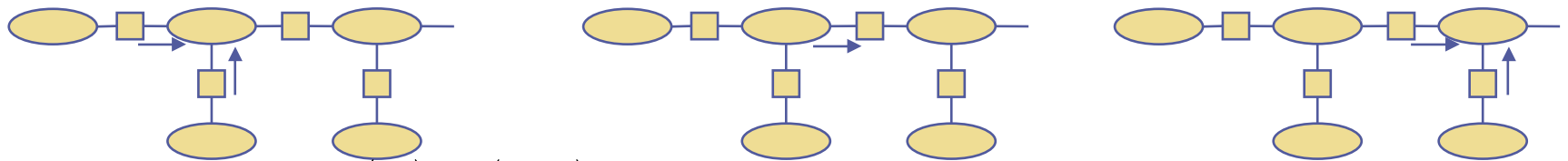
- Collect leftmost phi separator to the right:

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



# HMMs: Collect with Evidence

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



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

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

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

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

$$\phi^*(q_2) = \sum_{q_1} p(\bar{y}_0, \bar{y}_1, q_1) p(\bar{y}_2 | q_2) p(q_2 | q_1) = p(\bar{y}_0, \bar{y}_1, \bar{y}_2, q_2)$$

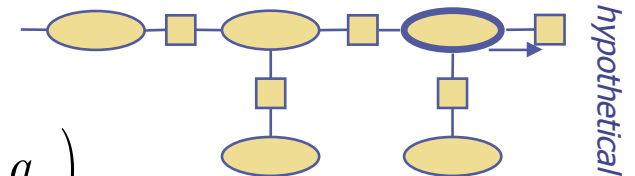
$$\phi^*(q_{t+1}) = \sum_{q_t} p(\bar{y}_0, \dots, \bar{y}_t, q_t) p(\bar{y}_{t+1} | q_{t+1}) p(q_{t+1} | q_t) = p(\bar{y}_0, \dots, \bar{y}_{t+1}, q_{t+1})$$

# HMMs: Evaluate with Evidence

- Say we are solving the first HMM problem:
  - 1) **Evaluate**: given  $y_0, \dots, y_T$  &  $\theta$  compute  $p(y_0, \dots, y_T | \theta)$
- If we want to compute the likelihood, we are already done!
- We really just need to do collect (not even distribute).
- From previous slide we had:

$$\phi^*(q_{t+1}) = \sum_{q_t} p(\bar{y}_0, \dots, \bar{y}_t, q_t) p(\bar{y}_{t+1} | q_{t+1}) p(q_{t+1} | q_t) = p(\bar{y}_0, \dots, \bar{y}_{t+1}, q_{t+1})$$

- Collect 'til root (rightmost node):  $\psi^*(q_{T-1}, q_T) = p(\bar{y}_0, \dots, \bar{y}_T, q_{T-1}, q_T)$   
its normalizer is  $p(\text{EVIDENCE})$ !



Or use hypothetical  $\phi^*(q_T) = p(\bar{y}_0, \dots, \bar{y}_T, q_T)$

- Can compute the likelihood just by marginalizing this phi

$$p(\bar{y}_0, \dots, \bar{y}_T) = \sum_{q_T} p(\bar{y}_0, \dots, \bar{y}_T, q_T) = \sum_{q_T} \phi^*(q_T)$$

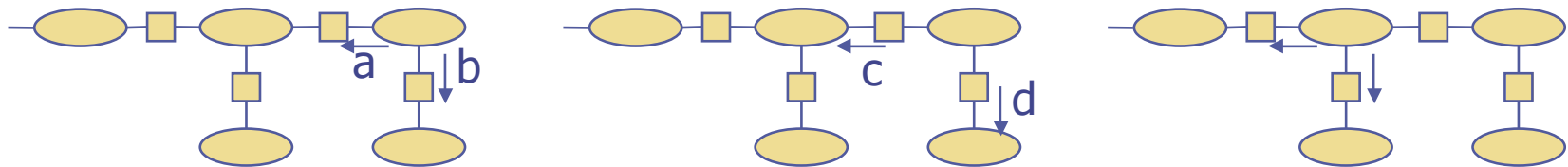
- So, adding up the entries in last  $\phi^*$  gives us the likelihood

# HMMs: Distribute with Evidence

- Back to collecting... say just finished collecting to the root with our last update formula:

$$\psi^*(q_{T-1}, q_T) = \frac{\phi^*(q_{T-1})}{1} \frac{\varsigma^*(q_T)}{1} \psi(q_{T-1}, q_T) = \phi^*(q_{T-1}) p(\bar{y}_T | q_T) \alpha_{q_{T-1}, q_T}$$

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



for  $t=T-1$  to 0

$$\begin{array}{ll} \text{a)} & \phi^{**}(q_t) = \sum_{q_{t+1}} \psi^{**}(q_t, q_{t+1}) \\ \text{b)} & \varsigma^{**}(q_{t+1}) = \sum_{q_t} \psi^{**}(q_t, q_{t+1}) \\ \text{c)} & \psi^{**}(q_t, q_{t+1}) = \frac{\phi^{**}(q_{t+1})}{\phi^*(q_{t+1})} \psi^*(q_t, q_{t+1}) \\ \text{d)} & \psi^{**}(y_t, q_t) = \frac{\varsigma^{**}(q_t)}{\varsigma^*(q_t)} \psi(y_t, q_t) \end{array}$$

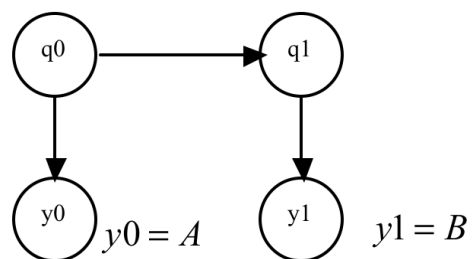
# HMM Example

You are given the parameters of a 2-state HMM. You observed the input sequence AB (from a 2-symbol alphabet A or B). In other words, you observe two symbols from your finite state machine, A and then B. Using the junction tree algorithm, evaluate the likelihood of this data  $p(y)$  given your HMM and its parameters. Also compute (for decoding) the individual marginals of the states after the evidence from this sequence is observed:  $p(q_0|y)$  and  $p(q_1|y)$ . The parameters for the HMM are provided below. They are the initial state prior  $p(q_0)$ , the state transition matrix given by  $p(q_t|q_{t-1})$ , and the emission matrix  $p(y_t|q_t)$ , respectively.

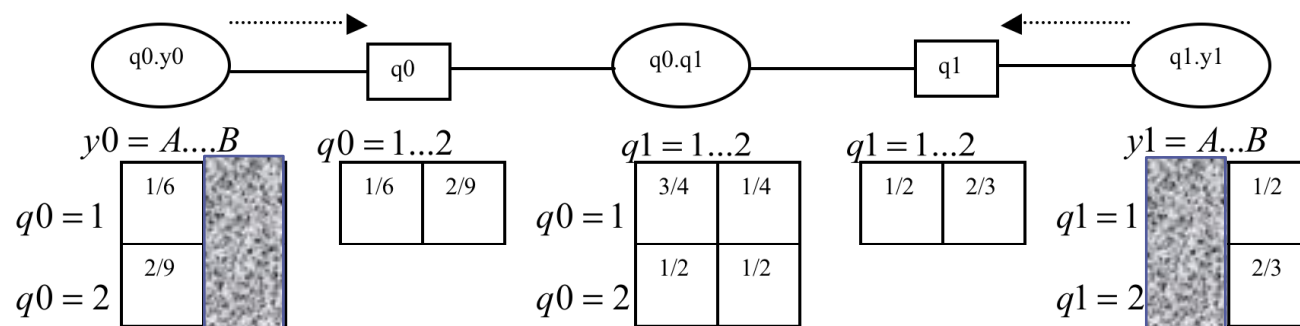
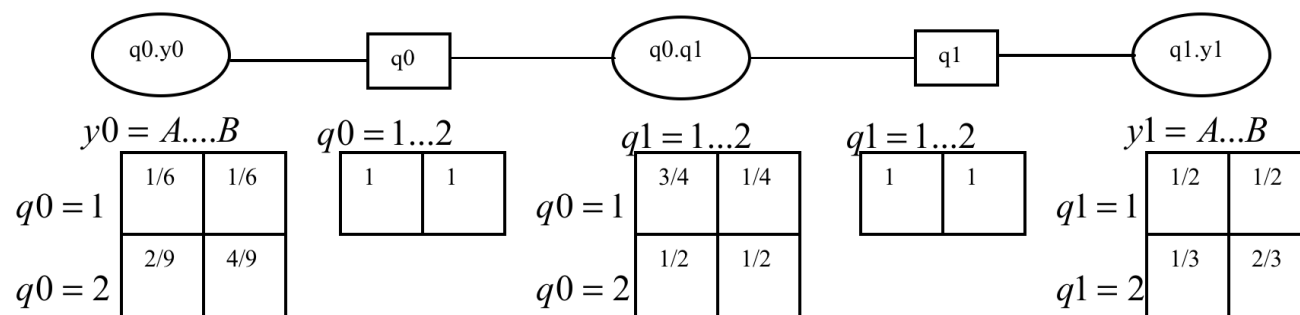
$$\pi = p(q_0) = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{bmatrix} 1/3 & 2/3 \end{bmatrix} \end{matrix}$$

$$a^T = p(q_t|q_{t-1}) = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{bmatrix} 3/4 & 1/2 \\ 1/4 & 1/2 \end{bmatrix} \end{matrix} \quad \eta^T = p(y_t|q_t) = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} A \\ B \end{matrix} & \begin{bmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{bmatrix} \end{matrix}$$

# HMM Example

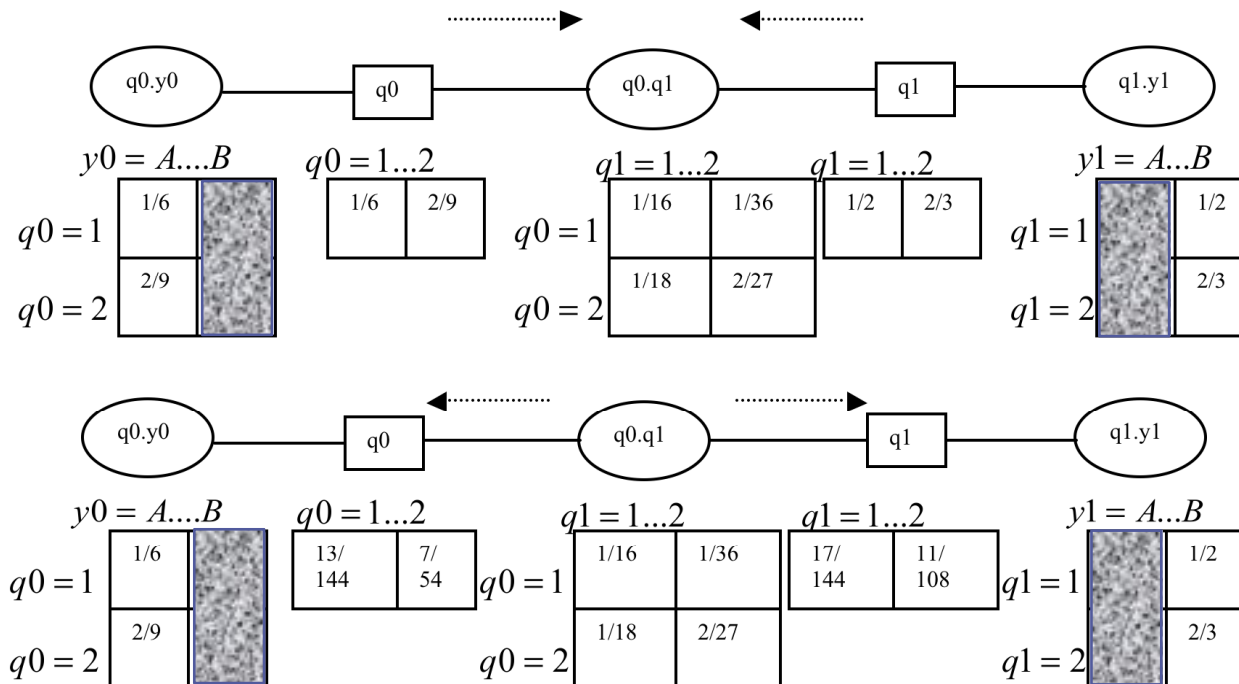


Initialized Junction Tree





# HMM Example



So the likelihood  $p(y) = \frac{13}{144} + \frac{7}{54} = \frac{1}{16} + \frac{1}{18} + \frac{1}{36} + \frac{2}{27} = \frac{17}{144} + \frac{11}{108} = \frac{95}{432} = 0.2199$

$$p(q_0 = 1 | y) = \frac{13/144}{13/144 + 7/54} = \frac{39}{95}, \quad p(q_0 = 2 | y) = \frac{7/54}{13/144 + 7/54} = \frac{56}{95}$$

$$p(q_1 = 1 | y) = \frac{17/144}{17/144 + 11/108} = \frac{51}{95}, \quad p(q_1 = 2 | y) = \frac{11/108}{17/144 + 11/108} = \frac{44}{95}$$

# HMMs: Marginals & MaxDecoding

- Now that JTA is finished, we have the following:

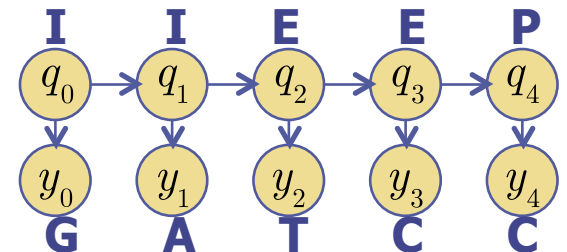
$$\phi^{**}(q_t) \propto p(q_t \mid \bar{y}_1, \dots, \bar{y}_T) \quad \varsigma^{**}(q_{t+1}) \propto p(q_{t+1} \mid \bar{y}_1, \dots, \bar{y}_T)$$

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

- The separators define a distribution over the hidden states
- This gives the probability the DNA symbol  $y_t$  was  $q_t = \{I, E, P\}$
- We've done 2) **Decode**: given  $y_0, \dots, y_T$  &  $\theta$  find  $p(q_0), \dots, p(q_T)$

- Can also do 2) **Decode**: given  $y_0, \dots, y_T$  &  $\theta$  find  $q_0, \dots, q_T$
- We can also decode to find the most likely path  $q_0 \dots q_T$
- Here, we use the ArgMax JTA algorithm
- Run JTA but replace sums with max
- Then, find biggest entry in separators:

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



# HMMs: EM Learning

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

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

**Introduce Lagrange & take derivatives**  $\longrightarrow$   $\sum_{i=1}^M \pi_i = 1 \quad \sum_{j=1}^M \alpha_{ij} = 1 \quad \sum_{j=1}^N \eta_{ij} = 1$

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

# HMMs: EM Learning

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

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

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

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

- **M-step** is maximizing as before:

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

- What are  $E\{\}$ 's?

# HMMs: EM Learning

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

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

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

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

- **M-step** is maximizing as before:

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

- What are  $E\{\}$ 's?

$$E_{p(x)}\{x^i\} = \sum_x p(x) x^i = \sum_x p(x) \delta(x - x^i) = p(x^i)$$

# HMMs: EM Learning

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

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

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

$$\begin{aligned} E\{l(\theta)\} &= E_{p(q_0, \dots, q_T | y)} \{\log p(q, y)\} = \sum_{q_0} \cdots \sum_{q_T} p(q | y) \log p(q, y) \\ &= E\left\{\sum_{i=1}^M q_0^i \log \pi_i + \sum_{t=1}^T \sum_{i,j=1}^M q_{t-1}^i q_t^j \log \alpha_{ij} + \sum_{t=0}^T \sum_{i=1}^M \sum_{j=1}^N q_t^i y_t^j \log \eta_{ij}\right\} \\ &= \sum_{i=1}^M E\{q_0^i\} \log \pi_i + \sum_{t=1}^T \sum_{i,j=1}^M E\{q_{t-1}^i q_t^j\} \log \alpha_{ij} + \sum_{t=0}^T \sum_{i=1}^M \sum_{j=1}^N E\{q_t^i\} y_t^j \log \eta_{ij} \end{aligned}$$

- **M-step** is maximizing as before:

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

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

- Our JTA  $\psi$  &  $\phi$  marginals! (JTA is the **E-Step** for given  $\theta$ )

$$E\{q_t^i q_{t+1}^j\} = p(q_t = i, q_{t+1} = j | \bar{y}) = \frac{\psi^{**}(q_t = i, q_{t+1} = j)}{\sum_{ij} \psi^{**}(q_t = i, q_{t+1} = j)} \quad E\{q_t^i\} = p(q_t = i | \bar{y}) = \frac{\phi^{**}(q_t = i)}{\sum_i \phi^{**}(q_t = i)}$$

# Thank you!

- So, to incomplete maximize likelihood with EM,
  - initialize parameters randomly,
  - Run Junction Tree Algorithm to get marginals
  - Use marginals over  $q$ 's in the maximum likelihood step
- Please complete course evaluation on courseworks
- Good luck with finals week and happy holidays!