Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2014

22 Parameter Estimation from Partial Observations

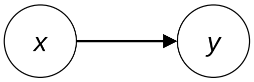
We've looked at learning parameters given graph structure for directed and undirected graphical model when all variables are observed. Today, we consider when we don't observe all the variables, focusing just on parameter estimation. A key issue we'll face is that parameters get coupled, complicating the estimation process.

22.1 Latent variables

There are numerous examples where we only observe variables partially and are required to learn the parameters associated with entire graphical model. For example, consider a classification task such as determining whether an email received is spam or ham. Unfortunately, email senders don't actually give a label for whether the email they sent is spam or ham. Thus, treating the label as a random variable, the label is hidden and we want to infer what the label is given observations, which for email classification could be features extracted from email content.

We give a sense of why latent variables could make estimation more difficult via the following example:

Example 1. Consider the following two-node graph.



With

$$\begin{aligned} p_{\mathsf{x},\mathsf{y}}(x, y; \theta) &= p_{\mathsf{y}}(y; \theta_{\mathsf{y}}) p_{\mathsf{x}|\mathsf{y}}(x|y; \theta_{\mathsf{x}|\mathsf{y}}), \\ &= \theta_{\mathsf{y}}(y) \theta_{\mathsf{x}|\mathsf{y}}(x|y). \end{aligned}$$

where

$$\theta_{\mathsf{y}} = [\theta_{\mathsf{y}}(y)]_{y \in \mathcal{Y}}, \qquad \theta_{\mathsf{x}|\mathsf{y}} = [\theta_{\mathsf{x}|\mathsf{y}}(x|y)]_{x \in \mathcal{X}, y \in \mathcal{Y}}.$$

Assuming that the model is fully observed, then as we've seen previously, the ML estimates come from empirical distribution matching:

$$\hat{\theta}_{\mathsf{y}}(y) = \hat{p}_{\mathsf{y}}(y), \qquad \hat{\theta}_{\mathsf{x}|\mathsf{y}}(x|y) = \frac{\hat{p}_{\mathsf{x},\mathsf{y}}(x, y)}{\hat{p}_{\mathsf{y}}(y)}.$$

Now consider if we instead only observed y. Then the log likelihood is given by

$$\ell(\theta; y) = \log \sum_{x} p_{\mathsf{x},\mathsf{y}}(x, y; \theta) = \log \sum_{x} \theta_{\mathsf{y}}(y) \theta_{\mathsf{x}|\mathsf{y}}(x|y).$$

for which we cannot push the log inside the summation. As a result, we cannot separate θ_y from $\theta_{x|y}$, so parameters θ_y and $\theta_{x|y}$ are said to "mix". Consequently, adjusting θ_y could affect $\theta_{x|y}$.

The example above imparts the issue of parameter coupling when dealing with latent variables. As a result, often we're left resorting to iterative algorithms that numerically compute ML estimates. For example, if the log likelihood is differentiable, then we can use gradient ascent. More generally, we can view ML estimation as a generic hill-climbing problem. But we haven't exploited the fact that we're dealing with a probability distribution, which is what we'll do next.

22.1.1 The Expectation-Maximization (EM) algorithm

Let $\mathsf{y} = (\mathsf{y}_1, \ldots, \mathsf{y}_N)$ be the set of observed variables and $\mathsf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_{N'})$ be the set of latent variables, where N and N' need not be the same. We'll refer to (y, x) as the complete data, where we'll assume that we know joint distribution $p_{\mathsf{y},\mathsf{x}}(\cdot, \cdot; \theta)$ with parameter θ. Given observation $\mathsf{y} = \mathbf{y}$, our goal is to find the ML estimate given by

$$\hat{\theta}^{\text{ML}} = \arg\max_{\theta} p_{\mathsf{y}}(\mathbf{y}; \theta) = \arg\max_{\theta} \underbrace{\log p_{\mathsf{y}}(\mathbf{y}; \theta)}_{\triangleq \ell(\theta; \mathbf{y})}.$$

We'll refer to the log likelihood $\ell(\theta; \mathbf{y})$ that we want to maximize as the *incomplete log likelihood*, whereas we'll call $\ell_c(\theta; \mathbf{y}, \mathbf{x}) = \log p_{\mathsf{y},\mathsf{x}}(\mathbf{y}, \mathbf{x}; \theta)$ the *complete log likelihood*, which would be the relevant log likelihood to maximize if we had observed the complete data.

The key idea behind the EM algorithm is the following. We will introduce a distribution q over hidden variables x that we get to choose. Notationally we'll write $q(\cdot|\mathbf{y})$ to indicate that we want q to depend on our (fixed) observations $\mathbf{y}$. Then

$$\begin{aligned}
\ell(\theta; \mathbf{y}) &= \log p_{\mathsf{y}}(\mathbf{y}; \theta) \\
&= \log \sum_{\mathbf{x}} p_{\mathsf{y},\mathsf{x}}(\mathbf{y}, \mathbf{x}; \theta) \\
&= \log \sum_{\mathbf{x}} q(\mathbf{x}|\mathbf{y}) \frac{p_{\mathsf{y},\mathsf{x}}(\mathbf{y}, \mathbf{x}; \theta)}{q(\mathbf{x}|\mathbf{y})} \\
&= \log \mathbb{E}_{q(\cdot|\mathbf{y})}\left[\frac{p_{\mathsf{y},\mathsf{x}}(\mathbf{y}, \mathbf{x}; \theta)}{q(\mathbf{x}|\mathbf{y})}\right] \\
&\geq \mathbb{E}_{q(\cdot|\mathbf{y})}\left[\log \frac{p_{\mathsf{y},\mathsf{x}}(\mathbf{y}, \mathbf{x}; \theta)}{q(\mathbf{x}|\mathbf{y})}\right] && \text{(Jensen's inequality)} \\
&= \sum_{\mathbf{x}} q(\mathbf{x}|\mathbf{y}) \log \frac{p_{\mathsf{y},\mathsf{x}}(\mathbf{y}, \mathbf{x}; \theta)}{q(\mathbf{x}|\mathbf{y})} \\
&\triangleq \mathcal{L}(q, \theta). && (1)
\end{aligned}$$

In particular, $\mathcal{L}(q,\theta)$ is easier to compute since we've pushed the log into the summation, and we were able to do so only because we introduced distribution q and used Jensen's inequality and concavity of the log function. So rather than maximizing $\ell(\theta;\mathbf{y})$, which is hard to compute, the EM algorithm will maximize lower bound $\mathcal{L}(q,\theta)$ by alternating between maximizing over q and then over θ:

E-step: $q^{(i+1)} = \arg\max_q \mathcal{L}(q,\theta^{(i)})$
M-step: $\theta^{i+1} = \arg\max_\theta \mathcal{L}(q^{(i+1)},\theta)$

We make a few remarks. First, in the M-step, by treating distribution q as fixed, we are solving

$$\begin{aligned}
\arg\max_\theta \mathcal{L}(q,\theta) &= \arg\max_\theta \mathbb{E}_{q(\cdot|\mathbf{y})}\left[\log \frac{p_{\mathsf{y},\mathsf{x}}(\mathbf{y},\mathbf{x};\theta)}{q(\mathbf{x}|\mathbf{y})}\right] \\
&= \arg\max_\theta \{\mathbb{E}_{q(\cdot|\mathbf{y})}[\log p_{\mathsf{y},\mathsf{x}}(\mathbf{y},\mathbf{x};\theta)] - \underbrace{\mathbb{E}_{q(\cdot|\mathbf{y})}[\log q(\mathbf{x}|\mathbf{y})]}_{\text{not dependent on }\theta}\} \\
&= \arg\max_\theta \underbrace{\mathbb{E}_{q(\cdot|\mathbf{y})}[\log p_{\mathsf{y},\mathsf{x}}(\mathbf{y},\mathbf{x};\theta)]}_{\substack{\triangleq \hat{\ell}_c^q(\theta;\mathbf{y}),\text{ the expected complete} \\ \text{log likelihood w.r.t. } q(\cdot|\mathbf{y})}} .
\end{aligned} \tag{2}$$

Note that $\hat{\ell}_c^q(\theta;\mathbf{y})$ is essentially an estimate of the incomplete log likelihood as we are taking the expectation over hidden variables with respect to $q(\cdot|\mathbf{y})$, effectively summing the hidden variables out but not through a marginalization. Also, note that its the expectation of the complete log likelihood, so deriving the optimal θ will turn out to be nearly identical to deriving an ML estimate for the complete log likelihood, where we will fill in expected counts related to hidden variables x.

Second, the E-step can be solved explicitly. Note that if we choose $q^{(i+1)}(\cdot|\mathbf{y}) = p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(i)})$, then

$$\begin{aligned}
\mathcal{L}(p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(i)}),\theta^{(i)}) &= \sum_{\mathbf{x}} p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)}) \log \frac{p_{\mathsf{y},\mathsf{x}}(\mathbf{y},\mathbf{x};\theta^{(i)})}{p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)})} \\
&= \sum_{\mathbf{x}} p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)}) \log \frac{p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)})p_{\mathsf{y}}(\mathbf{y};\theta^{(i)})}{p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)})} \\
&= \sum_{\mathbf{x}} p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)}) \log p_{\mathsf{y}}(\mathbf{y};\theta^{(i)}) \\
&= \underbrace{\left(\sum_{\mathbf{x}} p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)})\right)}_{=1} \log p_{\mathsf{y}}(\mathbf{y};\theta^{(i)}) \\
&= \log p_{\mathsf{y}}(\mathbf{y};\theta^{(i)}) \\
&= \ell(\theta^{(i)};\mathbf{y}).
\end{aligned}$$

So we have $\mathcal{L}(p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(i)}),\theta^{(i)}) = \ell(\theta^{(i)};\mathbf{y})$ while from inequality (1), we know that $\ell(\theta;\mathbf{y}) \geq \mathcal{L}(q,\theta)$ for all distributions q over random variable x, so it must be that choosing $q^{(i+1)}(\cdot|\mathbf{y}) = p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(i)})$ is optimal for solving the E-step.

By plugging the optimal choice of q from the E-step into the M-step and specifically equation (2), we see that we can combine both steps of an EM iteration into one update:

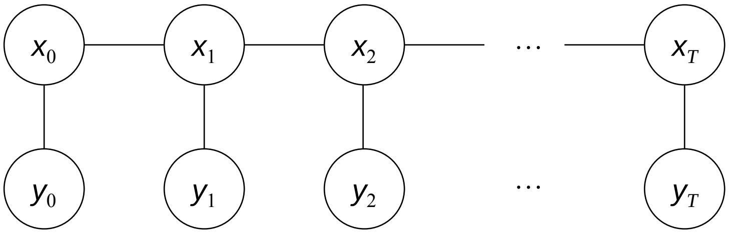
$$\begin{aligned}\theta^{(i+1)} &= \arg\max_{\theta} \mathbb{E}_{p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)})}[\log p_{\mathsf{y},\mathsf{x}}(\mathbf{y},\mathbf{x};\theta)] \\ &= \arg\max_{\theta} \mathbb{E}[\log p_{\mathsf{y},\mathsf{x}}(\mathbf{y},\mathsf{x};\theta)|\mathsf{y}=\mathbf{y};\theta^{(i)}]. \qquad (3)\end{aligned}$$

Hence, in the E-step, we can compute expectation $\mathbb{E}[\log p_{\mathsf{y},\mathsf{x}}(\mathbf{y},\mathsf{x};\theta)|\mathsf{y}=\mathbf{y};\theta^{(i)}]$ rather than computing out the full conditional distribution $p_{\mathsf{x}|\mathsf{y}}(\mathbf{x}|\mathbf{y};\theta^{(i)})$, and in the M-step, we maximize this expectation with respect to θ, justifying the name of the algorithm.

We end with two properties of the algorithm. First, the sequence of parameter estimates $\theta^{(i)}$ produced never decreases the likelihood and will reach a local maximum. Second, often the first few iterations will be the most useful and then the algorithm slows down. One way to accelerate convergence after the first few iterations is to switch to gradient ascent.

Example: Parameter estimation for hidden Markov models

Specializing to the case of HMM's, the EM algorithm is called the Baum-Welch algorithm (Baum *et al.* 1970; efficient implementation described by Welch in 2003). Consider a homogeneous HMM given by the diagram below.



We'll assume that $\mathsf{x}_i, \mathsf{y}_i$ take on values in $\{1,2,\ldots,M\}$. The model parameters are $\theta = \{\pi, A, \eta\}$ where:

- π is the initial state distribution ($\pi_i = p_{\mathsf{x}_0}(i)$)
- A is the transition matrix ($a_{ij} = p_{\mathsf{x}_{t+1}|\mathsf{x}_t}(j,i)$)
- η is the emission distribution ($\eta_{ij} = p_{\mathsf{y}_t|\mathsf{x}_t}(j|i)$)

We want an ML estimate of θ given only observation $\mathsf{y} = \mathbf{y}$.

We'll use the following data representation trick: Let $x_t^i = \mathbf{1}\{x_t = i\}$. Effectively, we represent state $x_t \in \{1,2,\ldots,M\}$ as a bit-string $(x_t^1, x_t^2, \ldots, x_t^M)$, where exactly one of the x_t^i's is 1 and the rest are 0. We'll do the same for y_t, letting $y_t^j = \mathbf{1}\{y_t = j\}$. This way of representing each state will prove extremely helpful in the sequel.

Before looking at the case where x is hidden, we first suppose that we observe the complete data where we have one observation for each x_i and one observation for each y_i. Then optimizing for the parameters involves maximizing the following:

$$
\begin{aligned}
&\log p_{\mathsf{x},\mathsf{y}}(\mathbf{x}, \mathbf{y}; \theta) \\
&= \log \left\{ \pi_{x_0} \prod_{t=0}^{T-1} a_{x_t, x_{t+1}} \prod_{t=0}^{T} \eta_{x_t, y_t} \right\} \\
&= \log \left\{ \left(\prod_{i=1}^{M} \pi_i^{\mathbf{1}\{x_0 = i\}} \right) \left(\prod_{t=0}^{T-1} \prod_{i=1}^{M} \prod_{j=1}^{M} a_{ij}^{\mathbf{1}\{x_t = i, x_{t+1} = j\}} \right) \left(\prod_{t=0}^{T} \prod_{i=1}^{M} \prod_{j=1}^{M} \eta_{ij}^{\mathbf{1}\{x_t = i, y_t = j\}} \right) \right\} \\
&= \log \left\{ \left(\prod_{i=1}^{M} \pi_i^{x_0^i} \right) \left(\prod_{t=0}^{T-1} \prod_{i=1}^{M} \prod_{j=1}^{M} a_{ij}^{x_t^i x_{t+1}^j} \right) \left(\prod_{t=0}^{T} \prod_{i=1}^{M} \prod_{j=1}^{M} \eta_{ij}^{x_t^i y_t^j} \right) \right\} \\
&= \sum_{i=1}^{M} x_0^i \log \pi_i + \sum_{t=0}^{T-1} \sum_{i=1}^{M} \sum_{j=1}^{M} x_t^i x_{t+1}^j \log a_{ij} + \sum_{t=0}^{T} \sum_{i=1}^{M} \sum_{j=1}^{M} x_t^i y_t^j \log \eta_{ij} \\
&= \sum_{i=1}^{M} \underbrace{x_0^i}_{\triangleq \zeta_i} \log \pi_i + \sum_{i=1}^{M} \sum_{j=1}^{M} \underbrace{\left[\sum_{t=0}^{T-1} x_t^i x_{t+1}^j \right]}_{\triangleq m_{ij}} \log a_{ij} + \sum_{i=1}^{M} \sum_{j=1}^{M} \underbrace{\left[\sum_{t=0}^{T} x_t^i y_t^j \right]}_{\triangleq n_{ij}} \log \eta_{ij}. \qquad (4)
\end{aligned}
$$

Note that m_{ij} is just the number of times we see state i followed by state j in the data. Similarly, n_{ij} is the number of times we see state i emit observation j.

By introducing Lagrange multipliers for constraints $\sum_{i=1}^{M} \pi_i = 1$, $\sum_{j=1}^{M} a_{ij} = 1$ for all i, and $\sum_{j=1}^{M} \eta_{ij} = 1$ for all i followed by setting a bunch of partial derivatives to 0, the ML estimates are given by:

$$
\widehat{\pi}_i = \zeta_i, \qquad \underbrace{\widehat{a}_{ij} = \frac{m_{ij}}{\sum_{k=1}^{M} m_{ik}}}_{\substack{\text{fraction of transitions} \\ \text{from } i \text{ that go to } j}}, \qquad \underbrace{\widehat{\eta}_{ij} = \frac{n_{ij}}{\sum_{k=1}^{M} n_{ik}}}_{\substack{\text{fraction of times state } i \\ \text{emits observation } j}} . \qquad (5)
$$

We now consider the case when x is hidden, so we'll use the EM algorithm. What we'll find is that our calculations above will not go to waste! Writing out the expectation to be maximized in an EM iteration, i.e. equation (3), we have

$$
\begin{aligned}
&\mathbb{E}_{p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}[\log p_{\mathsf{x},\mathsf{y}}(\mathbf{x},\mathbf{y};\theta)] \\
&= \mathbb{E}_{p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}\left[\sum_{i=1}^{M} \mathsf{x}_0^i \log \pi_i + \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\sum_{t=0}^{T-1} \mathsf{x}_t^i \mathsf{x}_{t+1}^j\right)\log a_{ij} + \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\sum_{t=0}^{T} \mathsf{x}_t^i y_t^j\right)\log \eta_{ij}\right] \\
&= \sum_{i=1}^{M} \mathbb{E}_{p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_0^i]\log\pi_i + \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\sum_{t=0}^{T-1}\mathbb{E}_{p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i \mathsf{x}_{t+1}^j]\right)\log a_{ij} \\
&\qquad + \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\sum_{t=0}^{T}\mathbb{E}_{p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i] y_t^j\right)\log\eta_{ij} \\
&= \sum_{i=1}^{M} \underbrace{\mathbb{E}_{p_{\mathsf{x}_0|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_0^i]}_{\zeta_i}\log\pi_i + \sum_{i=1}^{M}\sum_{j=1}^{M}\underbrace{\left(\sum_{t=0}^{T-1}\mathbb{E}_{p_{\mathsf{x}_t,\mathsf{x}_{t+1}|\mathsf{y}}(\cdot,\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i \mathsf{x}_{t+1}^j]\right)}_{m_{ij}}\log a_{ij} \\
&\qquad + \sum_{i=1}^{M}\sum_{j=1}^{M}\underbrace{\left(\sum_{t=0}^{T}\mathbb{E}_{p_{\mathsf{x}_t|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i] y_t^j\right)}_{n_{ij}}\log\eta_{ij}.
\end{aligned}
$$

Something magical happens: Note that using the sum-product algorithm with one forward and one backward pass, we can compute all node marginals $p_{\mathsf{x}_i|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})$ and edge marginals $p_{\mathsf{x}_i,\mathsf{x}_{i+1}|\mathsf{y}}(\cdot,\cdot|\mathbf{y};\theta^{(\ell)})$. Once we have these marginals, we can directly compute all the expectations $\mathbb{E}_{p_{\mathsf{x}_0|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_0^i]$ and $\mathbb{E}_{p_{\mathsf{x}_t,\mathsf{x}_{t+1}|\mathsf{y}}(\cdot,\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i \mathsf{x}_{t+1}^j]$; hence, we can compute all ζ_i, m_{ij}, and n_{ij}. Computing these expectations is precisely the E-step. As for the M-step, we can reuse our ML estimate formulas (5) from before except using our new values of ζ_i, m_{ij}, and n_{ij}.

For this problem, we can actually write down the expectations needed in terms of node and edge marginal probabilities. Using the fact that indictaor random variables are Bernoulli, we have:

- $\mathbb{E}_{p_{\mathsf{x}_t|\mathsf{y}}(\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i] = \mathbb{P}(\mathsf{x}_t = i|\mathsf{y} = \mathbf{y};\theta^{(\ell)})$
- $\mathbb{E}_{p_{\mathsf{x}_t,\mathsf{x}_{t+1}|\mathsf{y}}(\cdot,\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i \mathsf{x}_{t+1}^j] = \mathbb{P}(\mathsf{x}_t = i, \mathsf{x}_{t+1} = j|\mathsf{y} = \mathbf{y};\theta^{(\ell)})$

Then we can simplify our ML estimates as follows:

$$\widehat{\pi}_i^{(\ell+1)} = \zeta_i = \mathbb{E}_{p_{\mathsf{x}_0|\mathsf{y}}(\cdot|\mathbf{y};\widehat{\theta})}[\mathsf{x}_0^i] = \mathbb{P}(\mathsf{x}_0 = i|\mathsf{y} = \mathbf{y};\theta^{(\ell)}). \tag{6}$$

Next, note that the denominator of the ML estimate for $\widehat{a}_{ij}$ can be written as

$$\begin{aligned}
\sum_{k=1}^{M} m_{ik} &= \sum_{k=1}^{M}\sum_{t=0}^{T-1} \mathbb{E}_{p_{\mathsf{x}_t,\mathsf{x}_{t+1}|\mathbf{y}}(\cdot,\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i \mathsf{x}_{t+1}^k] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{p_{\mathsf{x}_t,\mathsf{x}_{t+1}|\mathbf{y}}(\cdot,\cdot|\mathbf{y};\theta^{(\ell)})}\left[\mathsf{x}_t^i \sum_{k=1}^{M} \mathsf{x}_{t+1}^k\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{p_{\mathsf{x}_t,\mathsf{x}_{t+1}|\mathbf{y}}(\cdot,\cdot|\mathbf{y};\theta^{(\ell)})}[\mathsf{x}_t^i] \qquad \text{(one bit is 1 and the rest are 0 for } \mathsf{x}_{t+1}) \\
&= \sum_{t=0}^{T-1} \mathbb{P}(\mathsf{x}_t = i|\mathbf{y} = \mathbf{y};\theta^{(\ell)}).
\end{aligned}$$

A similar result will hold for the denominator of the ML estimate for $\widehat{\eta}_{ij}$. Thus,

$$\widehat{a}_{ij}^{(\ell+1)} = \frac{m_{ij}}{\sum_{k=1}^{M} m_{ik}} = \frac{\sum_{t=0}^{T-1}\mathbb{P}(\mathsf{x}_t = i, \mathsf{x}_{t+1} = j|\mathbf{y} = \mathbf{y};\theta^{(\ell)})}{\sum_{t=0}^{T-1}\mathbb{P}(\mathsf{x}_t = i|\mathbf{y} = \mathbf{y};\theta^{(\ell)})} \tag{7}$$

$$\widehat{\eta}_{ij}^{(\ell+1)} = \frac{n_{ij}}{\sum_{k=1}^{M} n_{ik}} = \frac{\sum_{t=0}^{T}\mathbb{P}(\mathsf{x}_t = i, \mathsf{y}_t = j|\mathbf{y} = \mathbf{y};\theta^{(\ell)})}{\sum_{t=0}^{T}\mathbb{P}(\mathsf{x}_t = i|\mathbf{y} = \mathbf{y};\theta^{(\ell)})} \tag{8}$$

In summary, each iteration of the Baum-Welch algorithm does the following:

E-step: Given current parameter estimates $\theta^{(\ell)} = \{\widehat{\pi}^{(\ell)}, \widehat{A}^{(\ell)}, \widehat{\eta}^{(\ell)}\}$, run the sum-product algorithm to compute all node and edge marginals.
M-step: Using node and edge marginals, compute $\theta^{(\ell+1)} = \{\widehat{\pi}^{(\ell+1)}, \widehat{A}^{(\ell+1)}, \widehat{\eta}^{(\ell+1)}\}$ using equations (6), (7), and (8).

For the E-step, if we use the α-β forward-backward algorithm, edge marginals can be recovered using the following equation:

$$\begin{aligned}
&\mathbb{P}(\mathsf{x}_t = x_t, \mathsf{x}_{t+1} = x_{t+1}|\mathbf{y} = \mathbf{y}) \\
&\propto m_{t-1\to t}(x_t)\underbrace{\phi_t(x_t)}_{p(y_t|x_t)}\underbrace{\psi_{t,t+1}(x_t,x_{t+1})}_{p(x_{t+1}|x_t)}\underbrace{\phi_{t+1}(x_{t+1})}_{p(y_{t+1}|x_{t+1})}m_{t+2\to t+1}(x_{t+1}) \\
&= m_{t-1\to t}(x_t)p(y_t|x_t)p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})m_{t+2\to t+1}(x_{t+1}) \\
&= \left[\sum_{x_{t-1}} \alpha_{t-1}(x_{t-1})p(x_t|x_{t-1})\right] p(y_t|x_t)p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})\beta_{t+1}(x_{t+1}) \\
&\qquad \text{(recall how BP messages and } \alpha,\beta \text{ messages relate)} \\
&= \left[\sum_{x_{t-1}} \alpha_{t-1}(x_{t-1})p(x_t|x_{t-1})p(y_t|x_t)\right] p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})\beta_{t+1}(x_{t+1}) \\
&= \alpha_t(x_t)p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})\beta_{t+1}(x_{t+1}).
\end{aligned}$$

Hence, edge marginals at iteration ℓ are of the form:

$$\mathbb{P}(\mathsf{x}_t = i, \mathsf{x}_{t+1} = j | \mathbf{y} = \mathbf{y}; \theta^{(\ell)}) \propto \alpha_t(i) \widehat{a}_{ij}^{(\ell)} \widehat{\eta}_{j y_{t+1}}^{(\ell)} \beta_{t+1}(j). \tag{9}$$

We end our discussion of Baum-Welch with some key takeaways. First, note the interplay between inference and modeling: In the E-step, we're doing inference given our previous parameter estimates to obtain quantities that help us improve parameter estimates in the M-step.

Second, we were able to recycle our ML estimation calculations from the fully observed case. What we ended up introducing are essentially "soft-counts," e.g., the new m_{ij} is the expected number of times we see state i followed by state j.

Third, by linearity of expectation, the expected log likelihood for the HMM only ended up depending on expectations of nodes and edges, and as a result, we only needed to compute the node and edge marginals of distribution $q^{(\ell+1)} = p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y}; \theta^{(\ell)})$. We never needed to compute out the full joint distribution $p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathbf{y}; \theta^{(\ell)})$!