M step for HMM

We assume the E step computed distributions for

The degree each state explains each data point (analogous to GMM responsibilities).

$$\gamma(z_n) = p\left(z_n \middle| X, \theta^{(s)}\right)$$

The degree that the combination of a state, and a previous one explain two data points.

$$\xi(z_{n-1}, z_n) = p\left(z_{n-1}, z_n \middle| X, \theta^{(s)}\right)$$

"xi"

EM for HMM (sketch)

$$\log\left(p\left(X, Z \middle| \theta\right)\right) = \sum_{k=1}^{K} z_{1k} \log(\pi_k) + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} z_{n-1,j} z_{n,k} \log(A_{j,k}) + \sum_{n=1}^{N}\sum_{k}^{K} z_{n,k} \log\left(p\left(x_n \middle| \phi_k\right)\right)$$

By analogy with the GMM

$$Q\left(\theta^{(s+1)}, \theta^{(s)}\right) = \sum_{z} p\left(Z \middle| \theta^{(s)}\right) \log\left(X, Z \middle| \theta^{(s+1)}\right)$$

$$= \sum_{k=1}^{K} \gamma(z_{1,k}) \log(\pi_k) + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K} \xi(z_{n-1,j} z_{n,k}) \log(A_{j,k}) + \sum_{n=1}^{N}\sum_{k}^{K} \gamma(z_{nk}) \log\left(p\left(x_n \middle| z_n, \phi\right)\right)$$

EM for HMM (sketch)

Doing the maximization using Lagrange multipliers gives us

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{k'} \gamma(z_{1k'})}$$

Much like the GMM. Taking the partial derivative for π_k kills second and third terms.

$$A_{jk} = \frac{\sum_{n=2} \zeta(z_{n-1,j}, z_{nk})}{\sum_{k'}\sum_{n=2} \zeta(z_{n-1,j}, z_{nk'})}$$

EM for HMM (sketch)

The maximization of $p\left(x_n \middle| \phi\right)$ is exactly the same as the mixture model.

For example, if we have Gaussian emmisions, then

$$\mu_k = \frac{\sum_{n} x_n \gamma(z_{nk})}{\sum_{n} \gamma(z_{nk})}$$

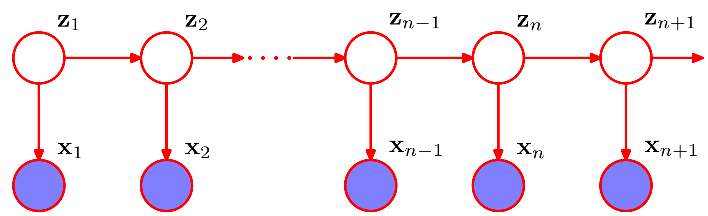
E step for EM for HMM

Computing the E step is a bit more involved.

Recall that in the mixture case it was easy because we only needed to consider the relative likelihood that each cluster independently explain the observations.

However, here the sequence also must play a role.

Graphical model for the E step



Note that our task is to compute marginal probabilities

Computing marginals in an HMM

Various names, flavors, notations, ...

Forward-Backward algorithm

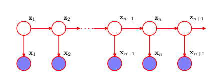
Alpha-beta algorithm

Sum-product for HMM

(Bishop also says "Baum Welch" but that is a synonym for the EM algorithm as whole).

Alpha-beta algorithm

$$\begin{aligned}
\gamma(z_n) &= p(z_n|X) \\
&= \frac{p(X|z_n)\,p(z_n)}{p(X)} \\
&= \frac{p(x_1,\ldots,x_n|z_n)\,p(x_{n+1},\ldots,x_N|z_n)\,p(z_n)}{p(X)} \\
&= \frac{p(x_1,\ldots,x_n,z_n)\,p(x_{n+1},\ldots,x_N|z_n)}{p(X)} \\
&= \frac{\alpha(z_n)\,\beta(z_n)}{p(X)}
\end{aligned}$$

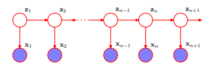


Where we define

$$\alpha(z_n) = p(x_1,\ldots,x_n,z_n)$$

$$\beta(z_n) = p(x_{n+1},\ldots,x_N|z_n)$$

Expressing alpha recursively



$$\alpha(z_n) = p(x_1,....,x_n,z_n)$$

$$= p(x_1,....,x_n|z_n)\,p(z_n) \qquad \text{(definition of "|")}$$

$$= p(x_n|z_n)\,p(x_1,....,x_{n-1}|z_n)\,p(z_n) \qquad \text{(conditional independence)}$$

$$= p(x_n|z_n)\,p(x_1,....,x_{n-1},z_n) \qquad \text{(definition of "|")}$$

$$= p(x_n|z_n)\sum_{z_{n-1}} p(x_1,....,x_{n-1},z_{n-1},z_n) \qquad \text{(marginal)}$$

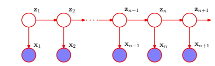
$$= p(x_n|z_n)\sum_{z_{n-1}} p(x_1,....,x_{n-1},z_n|z_{n-1})p(z_{n-1}) \qquad \text{(definition of "|")}$$

$$= p(x_n|z_n)\sum_{z_{n-1}} p(x_1,....,x_{n-1}|z_{n-1})p(z_n|z_{n-1})\,p(z_{n-1}) \qquad \text{(conditional independence)}$$

$$= p(x_n|z_n)\sum_{z_{n-1}} p(x_1,....,x_{n-1},z_{n-1})p(z_n|z_{n-1}) \qquad \text{(definition of "|")}$$

$$= p(x_n|z_n)\sum_{z_{n-1}} \alpha(z_{n-1})p(z_n|z_{n-1}) \qquad \text{(definition of } \alpha(z_n))$$

Expressing alpha recursively



$$\alpha(z_n) = p(x_n|z_n)\sum_{z_{n-1}} \alpha(z_{n-1})p(z_n|z_{n-1})$$

This is a recursive evaluation of alpha. So we can compute all of them easily if we know the first one, $\alpha(z_1)$.

$$\alpha(z_1) = p(x_1,z_1) \qquad \text{(we defined } \alpha(z_n) = p(x_1,....,x_n,z_n)\text{)}$$

$$= p(z_1)\,p(x_1|z_1) \qquad \text{(this is a K dimensional vector for fixed } x_1\text{)}$$

$$\alpha(z_1)_k = \pi_k\, p(x_1|\phi_k)$$

Alpha-beta algorithm

Similarly, we can derive a recurrence relation for beta

$$\beta(z_n) = \sum_{z_{n+1}} \beta(z_{n+1})p(x_{n+1}|z_{n+1})\,p(z_{n+1}|z_n)$$

Alpha-beta algorithm

The details for $\beta(z_n) = \sum_{z_{n+1}} \beta(z_{n+1})p(x_{n+1}|z_{n+1})\,p(z_{n+1}|z_n)$

$$\begin{aligned}
\beta(\mathbf{z}_n) &= p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N,\mathbf{z}_{n+1}|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_n,\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+1},\ldots,\mathbf{x}_N|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n) \\
&= \sum_{\mathbf{z}_{n+1}} p(\mathbf{x}_{n+2},\ldots,\mathbf{x}_N|\mathbf{z}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1})p(\mathbf{z}_{n+1}|\mathbf{z}_n).
\end{aligned}$$

Alpha-beta algorithm

Our recurrence relation for beta

$$\beta(z_n) = \sum_{z_{n+1}} \beta(z_{n+1}) p(x_{n+1}|z_{n+1}) p(z_{n+1}|z_n)$$

We can compute the betas if we know the last one.

$$\begin{aligned} p(z_N|X) &= \frac{\alpha(z_N)\beta(z_N)}{p(X)} \\ &= \frac{p(X,z_N)\beta(z_N)}{p(X)} \qquad \text{(we defined } \alpha(z_n) = p(x_1,....,x_n,z_n)\text{)} \\ &= p(z_N|X)\beta(z_N) \end{aligned}$$

So $\beta(z_N) = 1$

Alpha-beta algorithm

Given the alphas and betas, we can compute all the quantities we need for the E step.

$$\gamma(z_n) = p(z_n|X) = \frac{p(X|z_n)p(z_n)}{p(X)} = \frac{\alpha(z_n)\beta(z_n)}{p(X)} \qquad \text{(our definition)}$$

We know that $\sum_{z_n} \gamma(z_n) = 1$ (summing over the states)

so $\sum_{z_n} \frac{\alpha(z_n)\beta(z_n)}{p(X)} = 1$ (for all z_n)

and $p(X) = \sum_{z_n} \alpha(z_n)\beta(z_n)$ (for all z_n)

We do not need $p(X)$ for EM, but it is the likelihood which we want to monitor ($p(X) = p(X|\theta^{(s)})$).

Alpha-beta algorithm

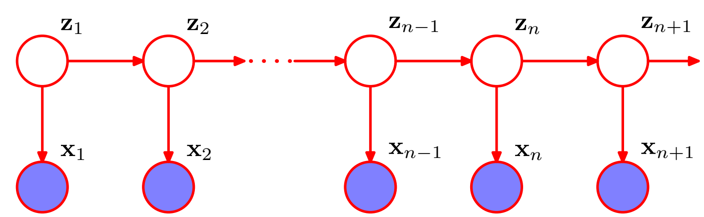
Given the alphas and betas, we can compute all the quantities we need for the E step.

$$\begin{aligned} \xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}) \\ &= \frac{p(\mathbf{X}|\mathbf{z}_{n-1}, \mathbf{z}_n)p(\mathbf{z}_{n-1}, \mathbf{z}_n)}{p(\mathbf{X})} \\ &= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}|\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})p(\mathbf{z}_{n-1})}{p(\mathbf{X})} \\ &= \frac{\alpha(\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})\beta(\mathbf{z}_n)}{p(\mathbf{X})} \end{aligned} \qquad (13.43)$$

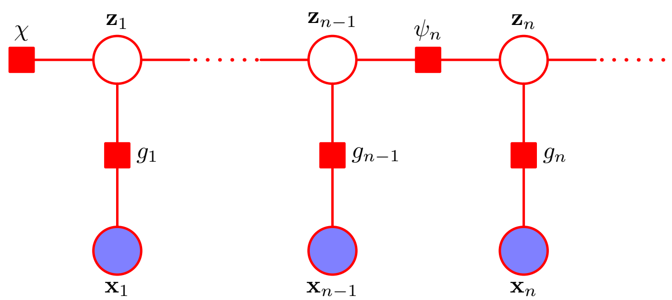
(in Bishop)

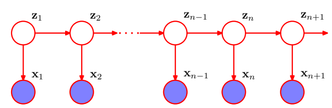
Computing marginals, version two

We can apply sum-product to our E step graph.



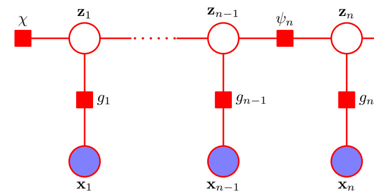
Factor graph





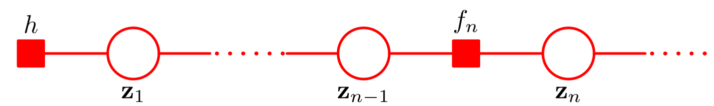
(Directed graph for reference)

Simplified factor graph

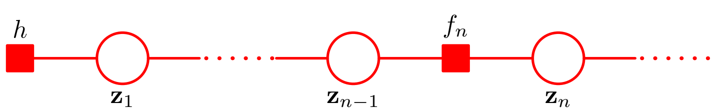


(Canonical factor graph from previous slide)

Since we condition on all the x's, we can simplify the graph by treating the emissions as constants, and putting them into the factors for the z's to get a simple chain.



Review of sum-product concepts



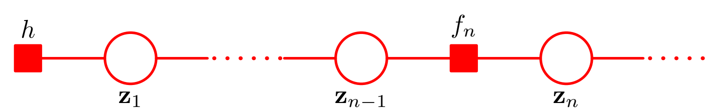
The marginal for each node is a product of the incoming messages.

This is analogous to setting up the marginal as a product of alpha and beta factors in the previous treatment.

Since we have a chain, this is just two messages, one coming from the left, the other from the right.

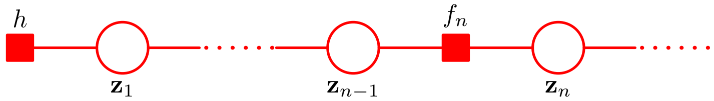
To compute all marginals, we pass the left and right messages from one end to the other.

Sum-product for HMM



$$h = p(z_1)\underbrace{p(x_1|z_1)}_{\text{extra for the nodes we pruned}}$$

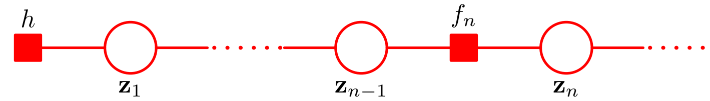
Sum-product for HMM



$$h = p(z_1)\,p(x_1|z_1)$$

$$f_n = p(z_n|z_{n-1})\underbrace{p(x_n|z_n)}_{\text{extra for pruned nodes}}$$

Sum-product for HMM

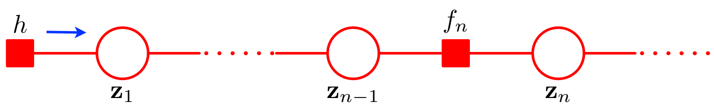


The nodes all have at most two links (it is a chain) so they just pass the incoming message to the outgoing link.

$$\text{i.e.,}\quad \mu_{f_n \to z_n}(z_n) = \mu_{f_n \to f_{n+1}}(z_n)$$

The nodes also (metaphorically) is where we think of the messages being stored if we are computing multiple marginals (which we are in this case).

Sum-product for HMM



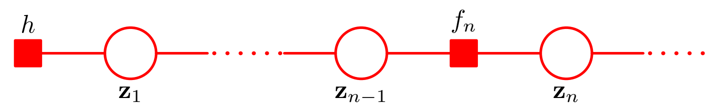
Factor node actions on left to right messages

$$\mu_{f_n \to f_{n+1}}(z_n) = \sum_{z_{n-1}} f_n(z_{n-1}, z_n)\,\mu_{f_{n-1} \to f_n}(z_{n-1})$$

The first message is

$$h = p(z_1)\,p(x_1|z_1) = p(x_1, z_1) = \alpha(z_1)$$

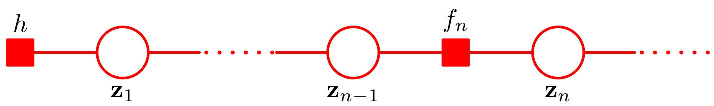
Sum-product for HMM



If we identify $\mu_{f_n \to f_{n+1}}(z_n) = \alpha(z_n)$

$$\begin{aligned}\alpha(z_n) &= \sum_{z_{n-1}} f_n(z_{n-1}, z_n)\,\alpha(z_{n-1}) \\ &= \sum_{z_{n-1}} p(z_n|z_{n-1})\,p(x_n|z_n)\,\alpha(z_{n-1}) \\ &= p(x_n|z_n)\sum_{z_{n-1}} p(z_n|z_{n-1})\,\alpha(z_{n-1}) \qquad \text{(as before)}\end{aligned}$$

Sum-product for HMM



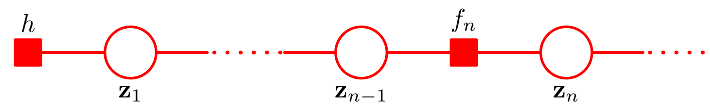
Factor node actions on right to left messages

$$\mu_{f_{n+1}\to f_n}(z_n) = \sum_{z_{n+1}} f_{n+1}(z_n, z_{n+1})\mu_{f_{n+2}\to f_{n+1}}(z_{n+1})$$

Identify $\beta(z_n) \equiv \mu_{f_{n+1}\to f_n}(z_n)$ to get

$$\beta(z_n) = \sum_{z_{n+1}} f_{n+1}(z_n, z_{n+1})\beta(z_{n+1})$$

Sum-product for HMM



Identify $\beta(z_n) \equiv \mu_{f_{n+1}\to f_n}(z_n)$ to get

$$\beta(z_n) = \sum_{z_{n+1}} f_{n+1}(z_n, z_{n+1})\beta(z_{n+1})$$

Recalling that $f_{n+1} = p(z_{n+1}|z_n)\,p(x_{n+1}|z_{n+1})$

$$\beta(z_n) = \sum_{z_{n+1}} p(z_{n+1}|z_n)p(x_{n+1}|z_{n+1})\beta(z_{n+1})$$

Sum-product for HMM

We have re-derived the alpha-beta version of forward-backward

Forward

$$\alpha(z_1) = p(z_1)\,p(x_1|z_1)$$

$$\alpha(z_n) = p(x_n|z_n)\sum_{z_{n-1}} p(z_n|z_{n-1})\alpha(z_{n-1})$$

Backward

$$\beta(z_N) = 1$$

$$\beta(z_n) = \sum_{z_{n+1}} p(z_{n+1}|z_n)p(x_{n+1}|z_{n+1})\beta(z_{n+1})$$

Sum-product for HMM

Also recall that the sum product enables easy computation of the normalizer (marginalizing everything), which corresponds to

$$p(X) = \sum_{z_n} \alpha(z_n)\beta(z_n)$$ (summing over clusters for any z_n)