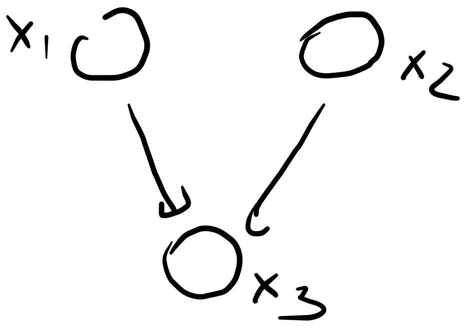
LEARNING BAYESIAN NETWORKS

$p(x) = \prod_i p(x_i \mid pa(x_i))$ $\qquad$ $p(x_i \mid pa(x_i), \vartheta)$, estimate by ML ϑ from data



$$P(x_1, x_2, x_3) = \underline{p(x_3 \mid x_1, x_2)}\, p(x_1)\, p(x_2)$$

$x_i \in \{0, 1\}$

$$p(x_3 = 1 \mid x_1 = 0, x_2 = 0) = \vartheta_{00} \longrightarrow \frac{\#(x_1 = 0, x_2 = 0, x_3 = 1)}{\#(x_1 = 0, x_2 = 0)} \quad ML.$$

$$p(x_3 = 1 \mid x_1 = 0, x_2 = 1) = \vartheta_{01}$$

$$p(x_3 = 1 \mid x_1 = 1, x_2 = 0) = \vartheta_{10}$$

$$p(x_3 = 1 \mid x_1 = 1, x_2 = 1) = \vartheta_{11}$$

What if some variables are LATENT (not-observed)?

X = observed

Z = latent

$p(x, z \mid \vartheta)$ and we want to estimate ϑ by ML.

$$p(x \mid \vartheta) = \sum_z p(x, z \mid \vartheta)$$

$x_1, \dots, x_N$ observation

$z_1, \dots, z_N$ latent states of observations

$$p(\underline{x}, \underline{z} \mid \vartheta) = \prod_n p(x_n, z_n \mid \vartheta) \qquad \log p(\underline{x}, \underline{z}) = \sum_n p(x_n, z_n)$$

$$\log p(\underline{x} \mid \vartheta) \neq \sum \log p(x_n \mid \vartheta)$$

ELBO

$p(x,z)$ x observed $x_1 \ldots x_N$ $p(x \mid \vartheta) = \sum_z p(x,z \mid \vartheta)$

z latent $z_1 \ldots z_N$

$\arg\max_{\vartheta} p(\underline{x} \mid \vartheta)$ ← function to optimize.

$p(x,z \mid \vartheta) = p(x \mid \vartheta)\underline{p(z \mid x, \vartheta)}$

VARIATIONAL APPROXIMATION $q(z)$ of $p(z \mid x, \vartheta)$

$$KL[q \| p] = KL[q(z) \| p(z \mid x, \vartheta)] = \mathbb{E}_{q(z)}\left[-\log \frac{p(z \mid x, \vartheta)}{q(z)}\right]$$

$$= -\sum_z q(z)\left[\log \frac{p(z \mid x, \vartheta)}{q(z)} + \log p(x \mid \vartheta) - \log p(x \mid \vartheta)\right]$$

$$= -\boxed{\sum_z q(z) \log \frac{p(x,z \mid \vartheta)}{q(z)}} + \log p(x \mid \vartheta)$$

$$f(q,\vartheta) = \sum_z q(z) \cdot \log \frac{p(x,z \mid \vartheta)}{q(z)}$$

$$\boxed{\log p(x \mid \vartheta) = \mathcal{L}(q, \vartheta) + KL\left[q(z) \,\|\, p(z \mid x, \vartheta)\right]}$$

$$\mathcal{L}(q, \vartheta) = \sum_{z} q(z) \cdot \log \frac{p(x, z \mid \vartheta)}{q(z)}$$

EVIDENCE LOWER BOUND (ELBO)

$$KL[q \,\|\, p] \geq 0 \Rightarrow \underbrace{\mathcal{L}(q, \vartheta) \leq \log(x \mid \vartheta)}$$

variational distribution

EXPECTATION MAXIMIZATION

$$\log p(\underline{x} \mid \vartheta) = \mathcal{L}(q, \vartheta) + KL\left[q(z) \,\|\, p(z \mid \underline{x}, \vartheta)\right]$$

$$\mathcal{L}(q, \vartheta) = \underbrace{\mathbb{E}_q\left[\log p(\underline{x}, z \mid \vartheta)\right]}_{\text{energy}} + \underbrace{\mathbb{E}_q\left[-\log q(z)\right]}_{H(q) \text{ entropy}} = \mathbb{E}_q\left[\log \frac{p(x, z \mid \vartheta)}{q(z)}\right]$$

GOAL $\vartheta_{ML} = \arg\max_{\vartheta} \log p(\underline{x} \mid \vartheta)$, intractable. THEN max $\mathcal{L}(q, \vartheta)$

E-step: maximise $\mathcal{L}(q, \vartheta)$ wrt $q(z)$, with ϑ FIXED to ϑ_{old}

q is maximized iff $KL\left[q \,\|\, p(z \mid x, \vartheta_{old})\right] = 0$

iff $q_{max}(z) = q(z) = p(\underline{z} \mid \underline{x}, \vartheta_{old})$

then compute $\mathbb{E}_{q_{new}}\left[\log p(\underline{x}, z \mid \vartheta)\right]$

$$p(\underline{z} \mid \underline{x}, \vartheta) = \prod_{i=1}^{N} p(z_i \mid x_i, \vartheta)$$

if $x_1 \ldots x_N$ are i.i.d.

$$p(\underline{z} \mid \underline{x}, \vartheta) = \frac{p(\underline{x}, \underline{z} \mid \vartheta)}{\sum_{z} p(\underline{x}, \underline{z} \mid \vartheta)} = \frac{\prod p(z_i, x_i \mid \vartheta)}{\sum_{\underline{z}} \prod p(z_i, x_i \mid \vartheta)} = \frac{\prod p(z_i, x_i \mid \vartheta)}{\prod \sum_{z_i} p(z_i, x_i \mid \vartheta)}$$

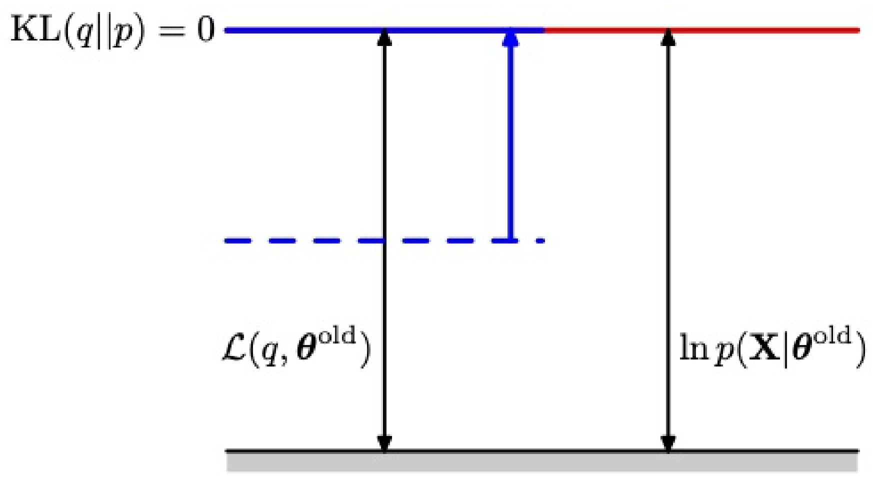
M-step maximize $\mathcal{L}(q, \theta)$ keeping q fixed to $q_{new} = p(\underline{z} \mid \underline{x}, \theta_{old})$

$\Updownarrow$

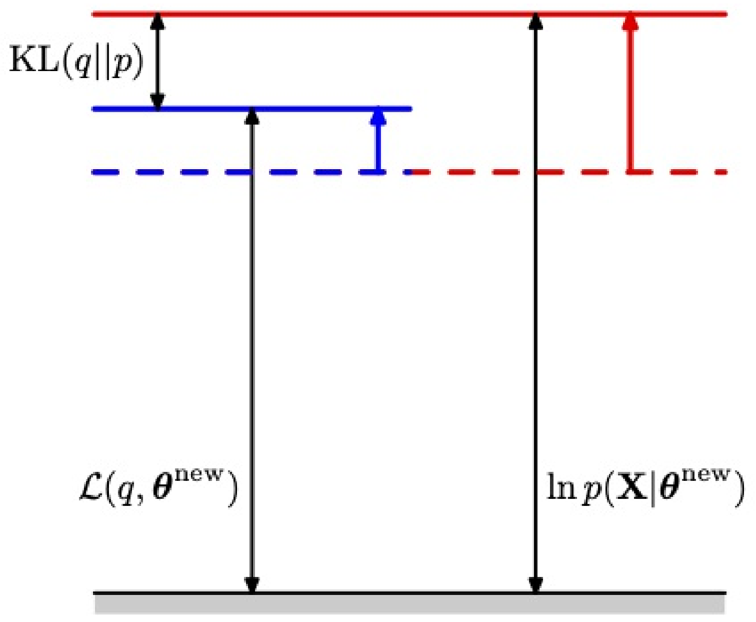
maximize $\mathbb{E}_{q_{new}}\left[\log p(\underline{z}, \underline{x} \mid \theta)\right]$

Then we have $\theta_{new} = \arg\max_{\theta} \mathbb{E}_{q_{new}}\left[\log p(\underline{x}, z \mid \theta)\right]$

Iterate E and M steps until convergence (when log-lik or $\|\theta_{old} - \theta_{new}\| < \epsilon$)

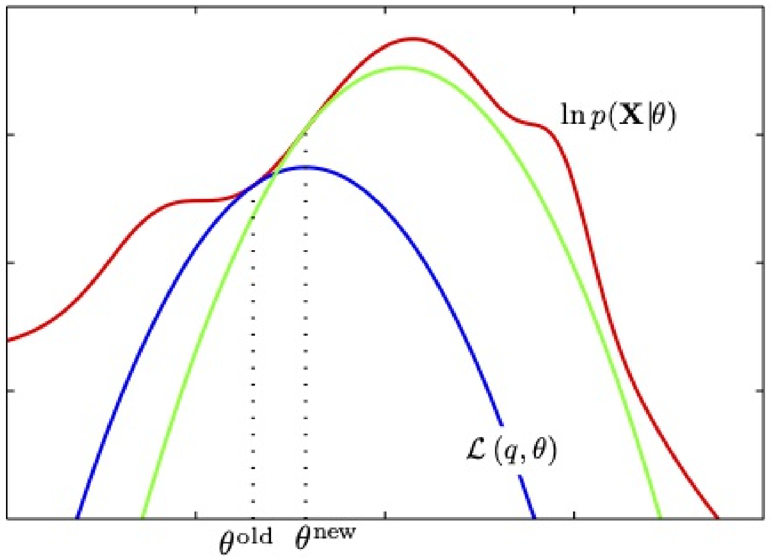


E-step



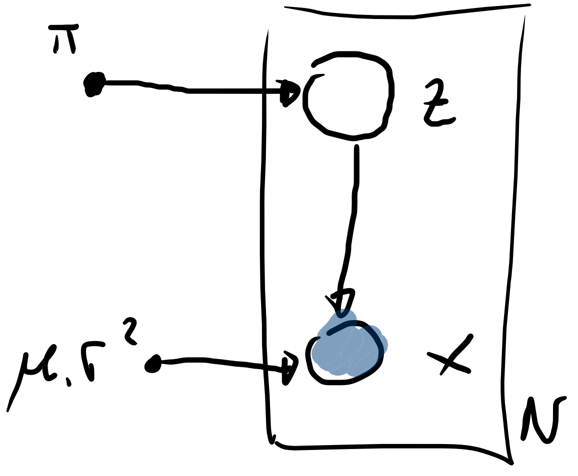
M-step

Figure 9.14 The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values. See the text for a full discussion.



EM converges to a local optimum of $\log p(\underline{x}|\theta)$

MIXTURE OF GAUSSIANS



z discrete: $z_1 \dots z_K$

$z_j \in \{0,1\}, \sum_j z_j = 1$

$\theta = (\pi, \mu, \sigma^2)$

$\underline{z} = (z_{ij}), \; i = 1 \to N, \; j = 1 \dots K$ not observed

$\underline{x} = x_i \quad , i = 1 \dots N$ observed

$$p(x, z \mid \theta) = \prod_{j=1}^{K} \pi_j^{z_j} \cdot \mathcal{N}(x \mid \mu_j, \sigma_j^2)^{z_j}$$

$$p(x \mid \theta) = \sum_{j=1}^{K} \pi_j \mathcal{N}(x \mid \mu_j, \sigma_j^2)$$

$$p(z = j \mid x, \theta) = \frac{\pi_j \mathcal{N}(x \mid \mu_j, \sigma_j^2)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(x \mid \mu_i, \sigma_i^2)}$$

$$p(z \mid \theta) = \prod_j \pi_j^{z_j}$$

$$p(\underline{z} \mid \underline{x}, \theta) \propto \prod_{n=1}^{N} \prod_{j=1}^{K} \pi_j^{z_{nj}} \mathcal{N}(x_n \mid \mu_j, \sigma_j^2)^{z_{nj}}$$

$$\mathbb{E}_{p(z|x)}[z_{nj}] = p(z_n = j \mid x_n, \vartheta) = \frac{\pi_j \mathcal{N}(x_n \mid \mu_j, \sigma_j^2)}{\sum_i \pi_i \mathcal{N}(x_n \mid \mu_i, \sigma_i^2)} = \underbrace{\gamma(z_{nj})}_{\text{RESPONSIBILITY}}$$

$$\log p(\underline{x}, \underline{z} \mid \vartheta) = \sum_{n=1}^{N} \sum_{j=1}^{K} z_{nj} \left[\log \pi_j + \log \mathcal{N}(x_n \mid \mu_j, \sigma_j^2) \right]$$

$$\mathbb{E}_{p(\underline{z} \mid \underline{x}, \vartheta)} \left[\log p(\underline{x}, \underline{z} \mid \vartheta) \right] = \sum_{n=1}^{N} \sum_{j=1}^{K} \underbrace{\mathbb{E}[z_{nj}]}_{\gamma(z_{nj})} \left[\log \underline{\pi_j} + \log \mathcal{N}(x_n \mid \underline{\mu_j}, \underline{\sigma_j^2}) \right]$$

E step.

$$\begin{cases} \mu_j^{new} = \frac{1}{N_j} \sum_n \gamma(z_{nj}) x_n \\ \Sigma_j^{new} = \frac{1}{N_j} \sum_n \gamma(z_{nj}) (x_n - \mu_j^{new})^T (x_n - \mu_j^{new}) \\ \pi_j^{new} = N_j / N , \quad N_j = \sum_{n=1}^{N} \gamma(z_{nj}) , \quad \sum_j N_j = N \end{cases}$$

M step

EM FOR BAYESIAN NETWORKS

$$p(x) = \prod_i p(x_i \mid pa(x_i), \vartheta_i) \qquad x = (v, z), \quad \vartheta = (\vartheta_i)_{i=1 \dots m}$$

$$p(x) = p(v, z \mid \vartheta)$$

$\rightarrow p(z \mid v = \hat{v}, \vartheta)$ for fixed ϑ

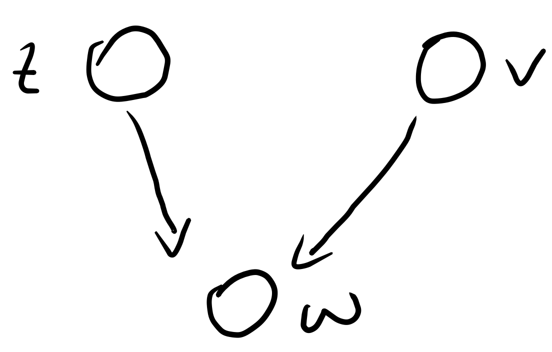
$\underline{v} = (v_1, \dots, v_N)$ observations of v ← E-step

$$q^n(z) = p(z \mid v_n, \vartheta) \rightsquigarrow \underline{q^n(x)} = p(z \mid v_n, \vartheta)\, \delta(v, v_n)$$

ENERGY for M-step

$$\sum_n \mathbb{E}_{q^n}\left[\log p(v_n, z_n \mid \vartheta)\right] = \sum_n \sum_i \mathbb{E}_{q^n}\left[\log p(x_i^n \mid pa(x_i^n), \vartheta_i)\right]$$

then optimize $\sum_n \mathbb{E}_{q^n}\left[\log p(x_i \mid pa(x_i), \vartheta_i)\right]$ over ϑ_i for each i



$, z, v, w \in \{0,1\}$

$p(z=1)=\vartheta_z \qquad x=(v,w,z)$

$p(v=1)=\vartheta_v$

$p(w=1 \mid z=a, v=b)=\underline{\vartheta_{wab}} \quad , a,b\in\{0,1\}$

$(v_1, w_1), \longrightarrow (v_N, w_N)$ observations

E-step

$$q^n(z)=p(z\mid v=v_n, w=w_n, \vartheta) \qquad q^n(x)=p(z\mid v=v_n, w=w_n, \vartheta)\,\delta(v,v_n)\,\delta(w,w_n)$$

$$\sum_n \mathbb{E}_{q^n}\Big[\log \underbrace{p(z^n\mid\vartheta_z)}_{\vartheta_z \text{ if } z^n=1}\Big] = \sum_n \log\vartheta_z\, q^n(z=1) + \log(1-\vartheta_z)\cdot q^n(z=0)$$

$$\vartheta_z = \frac{\sum_n q^n(z=1)}{\sum_n q^n(z=1)+\sum_n q^n(z=0)} = \frac{1}{N}\sum_n q^n(z=1)$$

$$\sum_n \mathbb{E}_{q^n}\left[\log p(w_n \mid z, v_n, \vartheta_w)\right]$$

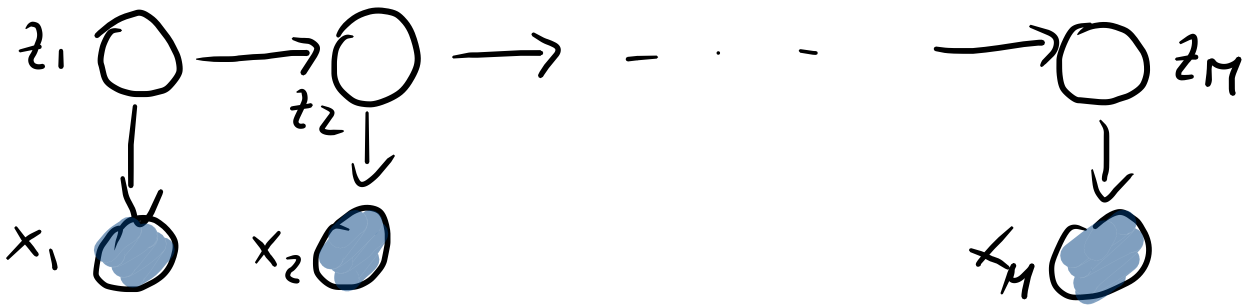
for $z=0$, $v=1$ (ϑ_{w01}):

$$\sum_{n:\, w_n=1,\, v_n=1} q^n(z=0)\cdot\log\vartheta_{w01} + \sum_{n:\, w_n=0,\, v_n=1} q^n(z=0)\log(1-\vartheta_{w01})$$

$$\Rightarrow \vartheta_{w01} = \frac{\sum_n \mathbb{I}(w_n=1)\,\mathbb{I}(v_n=1)\,q^n(z=0)}{\sum_n \mathbb{I}(w_n=1)\,\mathbb{I}(v_n=1)\,q^n(z=0) + \sum_n \mathbb{I}(w_n=0)\,\mathbb{I}(v_n=1)\,q^n(z=0)}$$

EM FOR HMM (Baum-Welch)

$z \in \{1 \ldots K\}$



$p(z_1 = i) = \pi_i$

$p(z_i = j \mid z_{i-1} = k) = A_{kj}$

$p(x_i \mid z_i = k) = p(x_i \mid \phi_k)$

$\theta = (\pi, A, \phi)$

$\underline{X} = x^1, \ldots, x^N$

$x^1 = (x_1^1, \ldots, x_M^1)$

E step $\quad q^n(z) = \underbrace{p(z \mid x^n, \theta)}_{\text{message passing.}} \quad \forall n$

$$\boxed{\pi_k = \frac{\sum_n q^n(z_{1k})}{\sum_j \sum_n q^n(z_{1j})}}$$

M step

$$E(\theta) = \sum_{n=1}^{N} \Big[\sum_{k=1}^{K} q^n(z_{1k}) \ln \pi_k + \sum_{i=2}^{M} \sum_{j,k=1}^{K} q^n(z_{i-1,j}, z_{ik}) \ln A_{jk} + \sum_{i=1}^{M} \sum_{k=1}^{K} q^n(z_{ik}) \ln p(x_i^n \mid \phi_k) \Big]$$