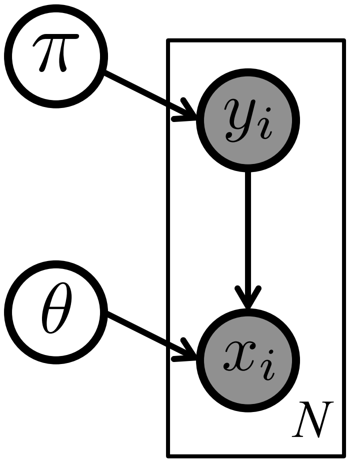
Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

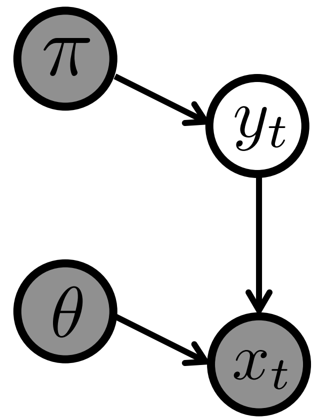
Lecture 9:
Expectation Maximization (EM) Algorithm,
Learning in Undirected Graphical Models

Some figures courtesy Michael Jordan's draft textbook,
An Introduction to Probabilistic Graphical Models

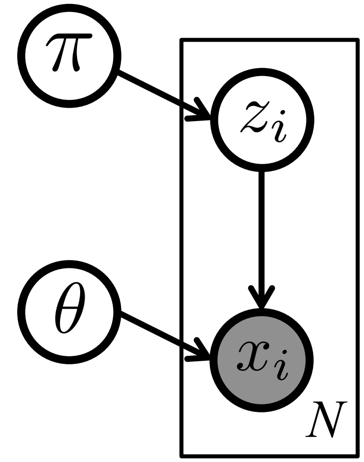
Expectation Maximization (EM)



Supervised Training



Supervised Testing



Unsupervised Learning

$\pi, \theta \longrightarrow$ *parameters (shared across observations)*

$z_1, \ldots, z_N \longrightarrow$ *hidden data (unique to particular instances)*

- **Initialization:** Randomly select starting parameters
- **E-Step:** Given parameters, find posterior of hidden data
  - Equivalent to test inference of full posterior distribution
- **M-Step:** Given posterior distributions, find likely parameters
  - Distinct from supervised ML/MAP, but often still tractable
- **Iteration:** Alternate E-step & M-step until convergence

EM as Lower Bound Maximization

$$\ln p(x \mid \theta) = \ln \left(\sum_z p(x, z \mid \theta) \right)$$

$$\ln p(x \mid \theta) \geq \sum_z q(z) \ln \left(\frac{p(x, z \mid \theta)}{q(z)} \right)$$

$$\ln p(x \mid \theta) \geq \sum_z q(z) \ln p(x, z \mid \theta) - \sum_z q(z) \ln q(z) \triangleq \mathcal{L}(q, \theta)$$

- **Initialization:** Randomly select starting parameters $\theta^{(0)}$
- **E-Step:** Given parameters, find posterior of hidden data

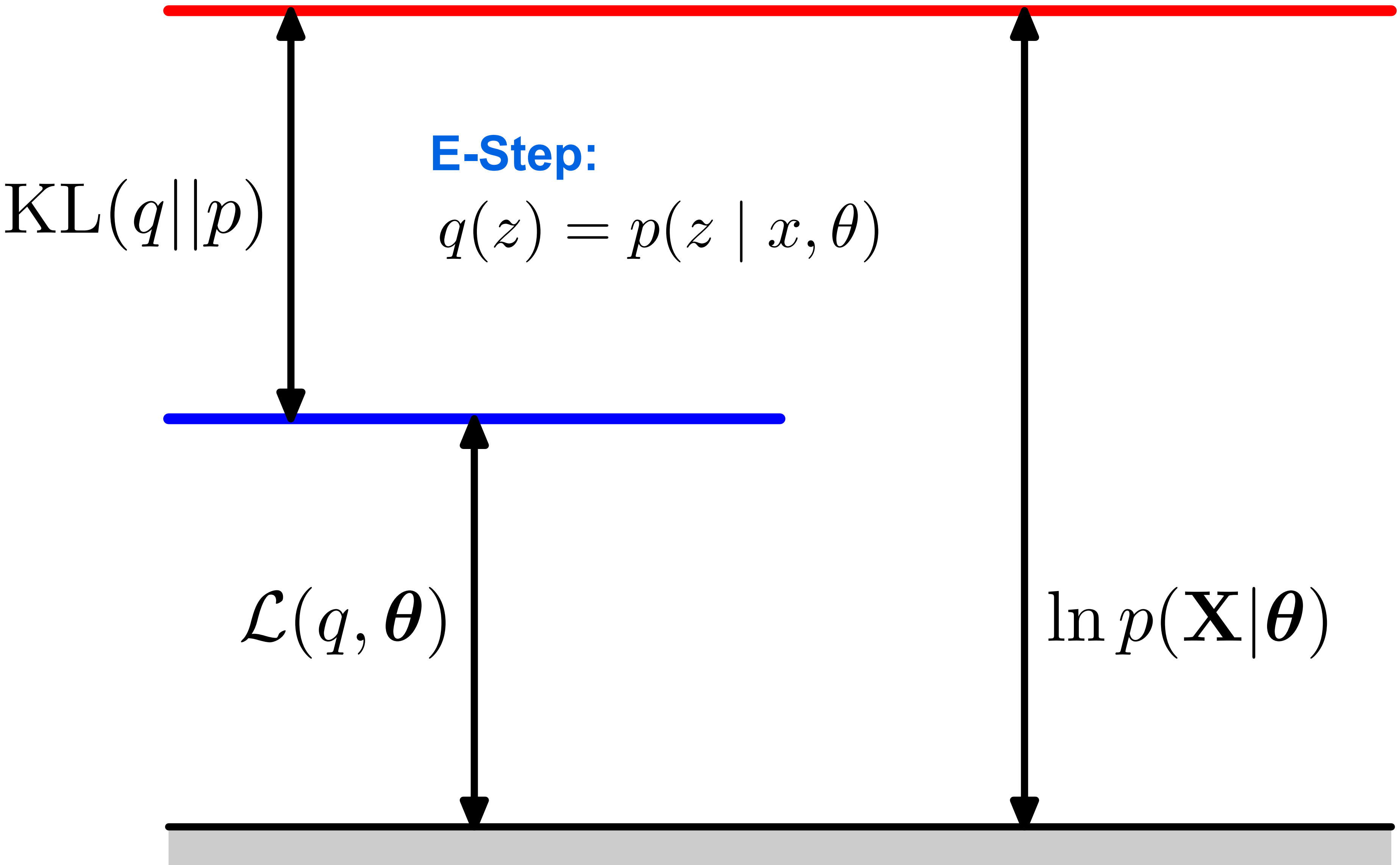
$$q^{(t)} = \arg\max_q \mathcal{L}(q, \theta^{(t-1)})$$

- **M-Step:** Given posterior distributions, find likely parameters

$$\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta)$$

- **Iteration:** Alternate E-step & M-step until convergence

Lower Bounds on Marginal Likelihood



C. Bishop, Pattern Recognition & Machine Learning

EM: Expectation Step

$$\ln p(x \mid \theta) \geq \sum_z q(z) \ln p(x, z \mid \theta) - \sum_z q(z) \ln q(z) \triangleq \mathcal{L}(q, \theta)$$

$$q^{(t)} = \arg\max_q \mathcal{L}(q, \theta^{(t-1)})$$

- General solution, for any probabilistic model:

$$q^{(t)}(z) = p(z \mid x, \theta^{(t-1)})$$

posterior distribution given current parameters

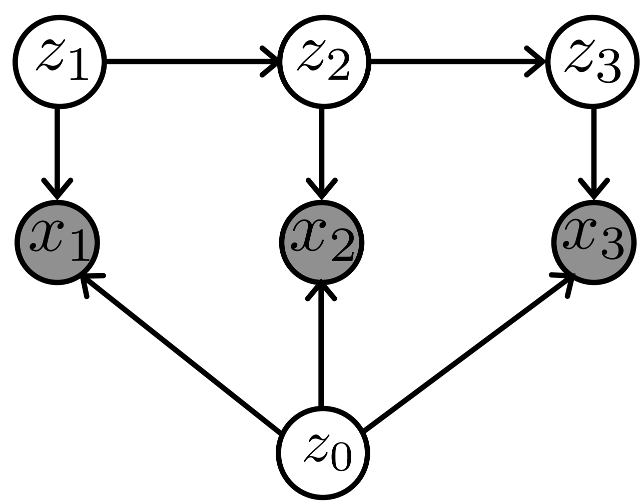
- For a directed graphical model:

$\theta \longrightarrow$ *fixes conditional distributions of every child node, given parents*

$x \longrightarrow$ *observed nodes (training data)*

$z \longrightarrow$ *unobserved nodes (hidden data)*

Inference: Find summary statistics of posterior needed for following M-step



EM: Maximization Step

$$\ln p(x \mid \theta) \geq \sum_z q(z) \ln p(x, z \mid \theta) - \sum_z q(z) \ln q(z) \triangleq \mathcal{L}(q, \theta)$$

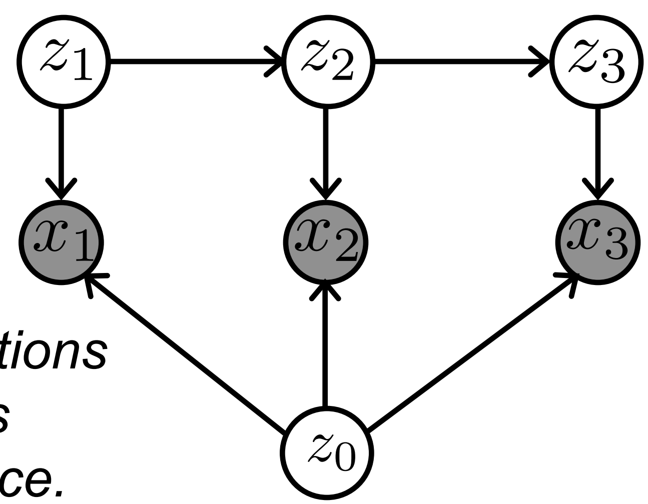
$$\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta) = \arg\max_\theta \sum_z q(z) \ln p(x, z \mid \theta)$$

- Recall directed graphical model factorization: $y = \{x, z\}$

$$\log p(y \mid \theta) = \sum_{n=1}^{N} \sum_{i \in \mathcal{V}} \log p(y_{i,n} \mid y_{\Gamma(i),n}, \theta_i) = \sum_{i \in \mathcal{V}} \left[\sum_{n=1}^{N} \log p(y_{i,n} \mid y_{\Gamma(i),n}, \theta_i) \right]$$

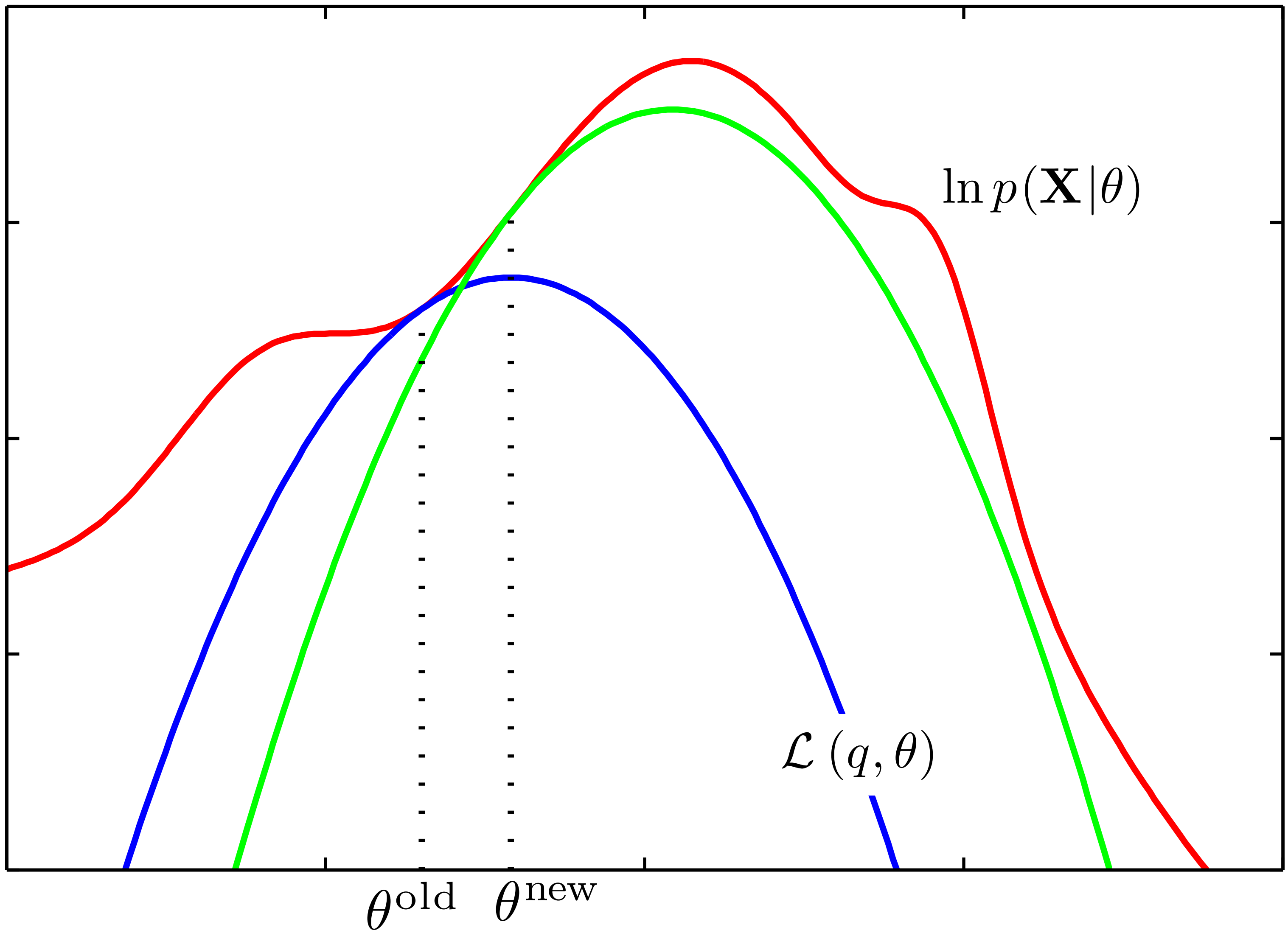
$$\mathbb{E}_q[\log p(y \mid \theta)] = \sum_{i \in \mathcal{V}} \left[\sum_{n=1}^{N} \mathbb{E}_{q_i}[\log p(y_{i,n} \mid y_{\Gamma(i),n}, \theta_i)] \right]$$

$$q_i(y_i, y_{\Gamma(i)}) = p(y_i, y_{\Gamma(i)} \mid \theta_i^{\text{old}}, x)$$



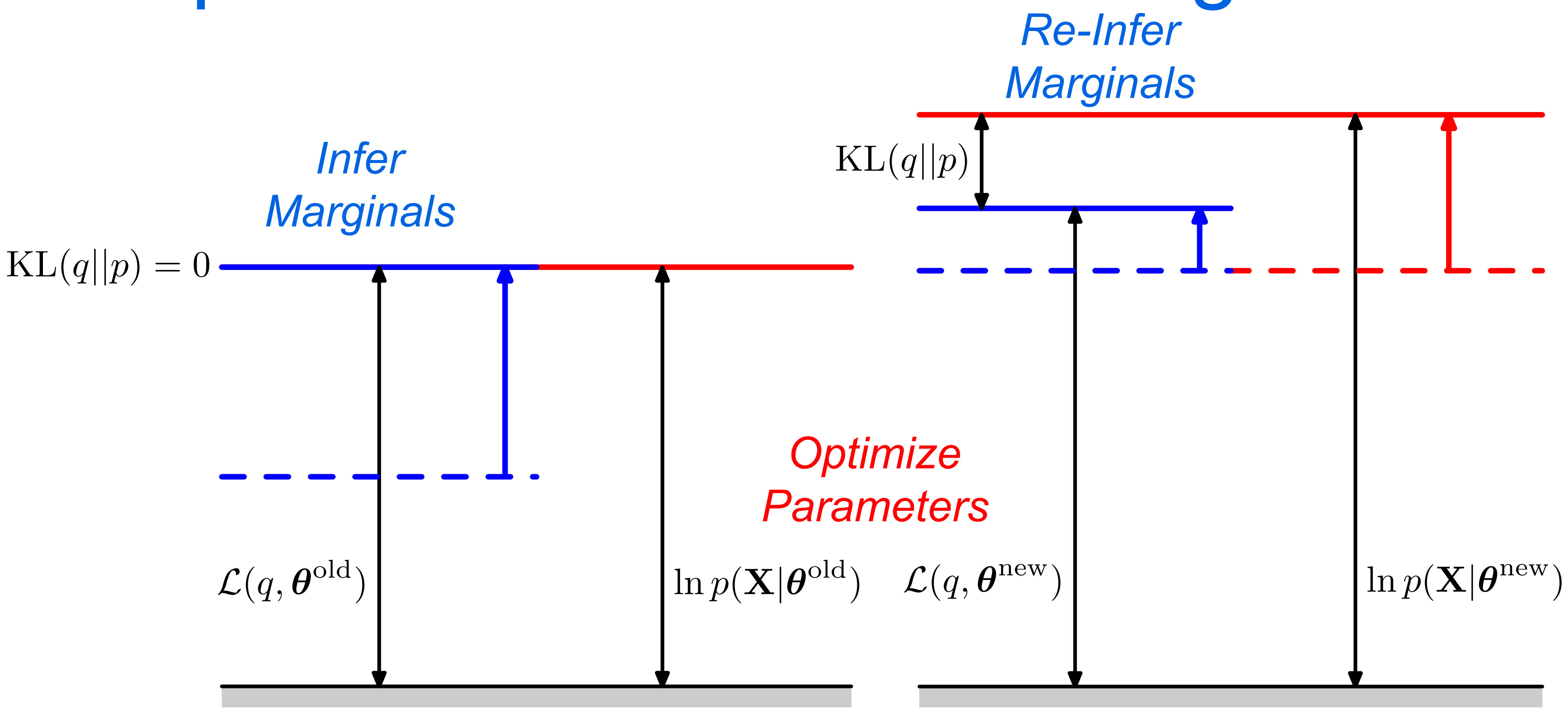
From E-step, only require posterior marginal distributions of each node and its parents, given observations (which have probability one) for that training instance.

EM: A Sequence of Lower Bounds



C. Bishop, Pattern Recognition & Machine Learning

Expectation Maximization Algorithm



E Step: *Optimize distribution on hidden variables given parameters*

M Step: *Optimize parameters given distribution on hidden variables*

C. Bishop, Pattern Recognition & Machine Learning

M-Step for Exponential Families

- Exponential Family: $p(x, z \mid \theta) = \exp\{\theta^T \phi(x, z) - A(\theta)\}$
- E-step Produces:

$$q(z_i), \;\; i = 1, \ldots, N \qquad \pi_{ik} \triangleq q(z_i = k)$$

- M-step Objective:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \mathbb{E}_{q_i}[\log p(x_i, z_i \mid \theta)] = \sum_{i=1}^{N} \mathbb{E}_{q_i}[\theta^T \phi(x_i, z_i) - A(\theta)]$$

- Taking gradient shows that optimum parameters satisfy

$$\mathbb{E}_{\hat{\theta}}[\phi(x, z)] = \frac{1}{N} \sum_{i=1}^{N} \sum_{k} \pi_{ik} \phi(x_i, k)$$

- As in basic mixture models, solution always matches moments:
  - For observed variables, empirical distribution of data
  - For hidden variables, weighted distribution from E-step
- In directed graphical models, apply to each local conditional…

EM for MAP Estimation

Up to a constant independent of $\theta, \; \ln p(\theta \mid x) =$

$$\ln p(\theta) + \ln p(x \mid \theta) = \ln p(\theta) + \ln\left(\sum_z p(x, z \mid \theta)\right)$$

$$\geq \ln p(\theta) + \sum_z q(z) \ln p(x, z \mid \theta) - \sum_z q(z) \ln q(z) \triangleq \mathcal{L}(q, \theta)$$

- **Initialization:** Randomly select starting parameters $\theta^{(0)}$
- **E-Step:** Given parameters, find posterior of hidden data

$$q^{(t)} = \arg\max_q \mathcal{L}(q, \theta^{(t-1)}) \qquad q^{(t)}(z) = p(z \mid x, \theta^{(t-1)})$$

- **M-Step:** Given posterior distributions, find likely parameters

$$\theta^{(t)} = \arg\max_\theta \mathcal{L}(q^{(t)}, \theta)$$

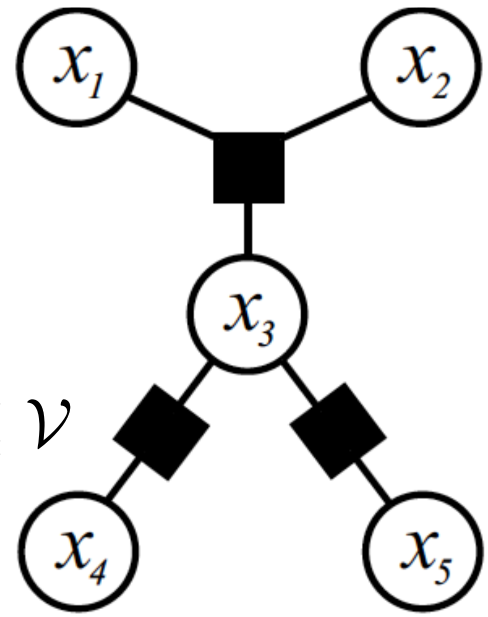
Objective is sum of prior and weighted likelihood

- **Iteration:** Alternate E-step & M-step until convergence

Undirected Graphical Models

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \prod_{f \in \mathcal{F}} \psi_f(x_f \mid \theta_f)$$

$$Z(\theta) = \sum_x \prod_{f \in \mathcal{F}} \psi_f(x_f \mid \theta_f)$$



$\mathcal{F} \longrightarrow$ set of hyperedges linking subsets of nodes $f \subseteq \mathcal{V}$

$\mathcal{V} \longrightarrow$ set of N nodes or vertices, $\{1, 2, \ldots, N\}$

- Assume an exponential family representation of each factor:

$$p(x \mid \theta) = \exp\left\{ \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_f) - A(\theta) \right\}$$

$$\psi_f(x_f \mid \theta_f) = \exp\{\theta_f^T \phi_f(x_f)\} \qquad A(\theta) = \log Z(\theta)$$

- Partition function *globally* couples the local factor parameters

Learning for Undirected Models

- Undirected graph encodes dependencies within a single training example:

$$p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} \frac{1}{Z(\theta)} \prod_{f \in \mathcal{F}} \psi_f(x_{f,n} \mid \theta_f) \qquad \mathcal{D} = \{x_{\mathcal{V},1}, \ldots, x_{\mathcal{V},N}\}$$

- Given N independent, identically distributed, completely observed samples:

$$\log p(\mathcal{D} \mid \theta) = \left[\sum_{n=1}^{N} \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_{f,n}) \right] - N A(\theta)$$

$$p(x \mid \theta) = \exp\left\{ \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_f) - A(\theta) \right\}$$

$$\psi_f(x_f \mid \theta_f) = \exp\{\theta_f^T \phi_f(x_f)\} \qquad A(\theta) = \log Z(\theta)$$

- Partition function *globally* couples the local factor parameters

Learning for Undirected Models

- Undirected graph encodes dependencies within a single training example:

$$p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} \frac{1}{Z(\theta)} \prod_{f \in \mathcal{F}} \psi_f(x_{f,n} \mid \theta_f) \qquad \mathcal{D} = \{x_{\mathcal{V},1}, \dots, x_{\mathcal{V},N}\}$$

- Given N independent, identically distributed, completely observed samples:

$$\log p(\mathcal{D} \mid \theta) = \left[\sum_{n=1}^{N} \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_{f,n}) \right] - N A(\theta)$$

- Take gradient with respect to parameters for a single factor:

$$\nabla_{\theta_f} \log p(\mathcal{D} \mid \theta) = \left[\sum_{n=1}^{N} \phi_f(x_{f,n}) \right] - N \mathbb{E}_\theta[\phi_f(x_f)]$$

- Must be able to compute *marginal distributions* for factors in current model:
  - Tractable for tree-structured factor graphs via sum-product
  - What about general factor graphs or undirected graphs?