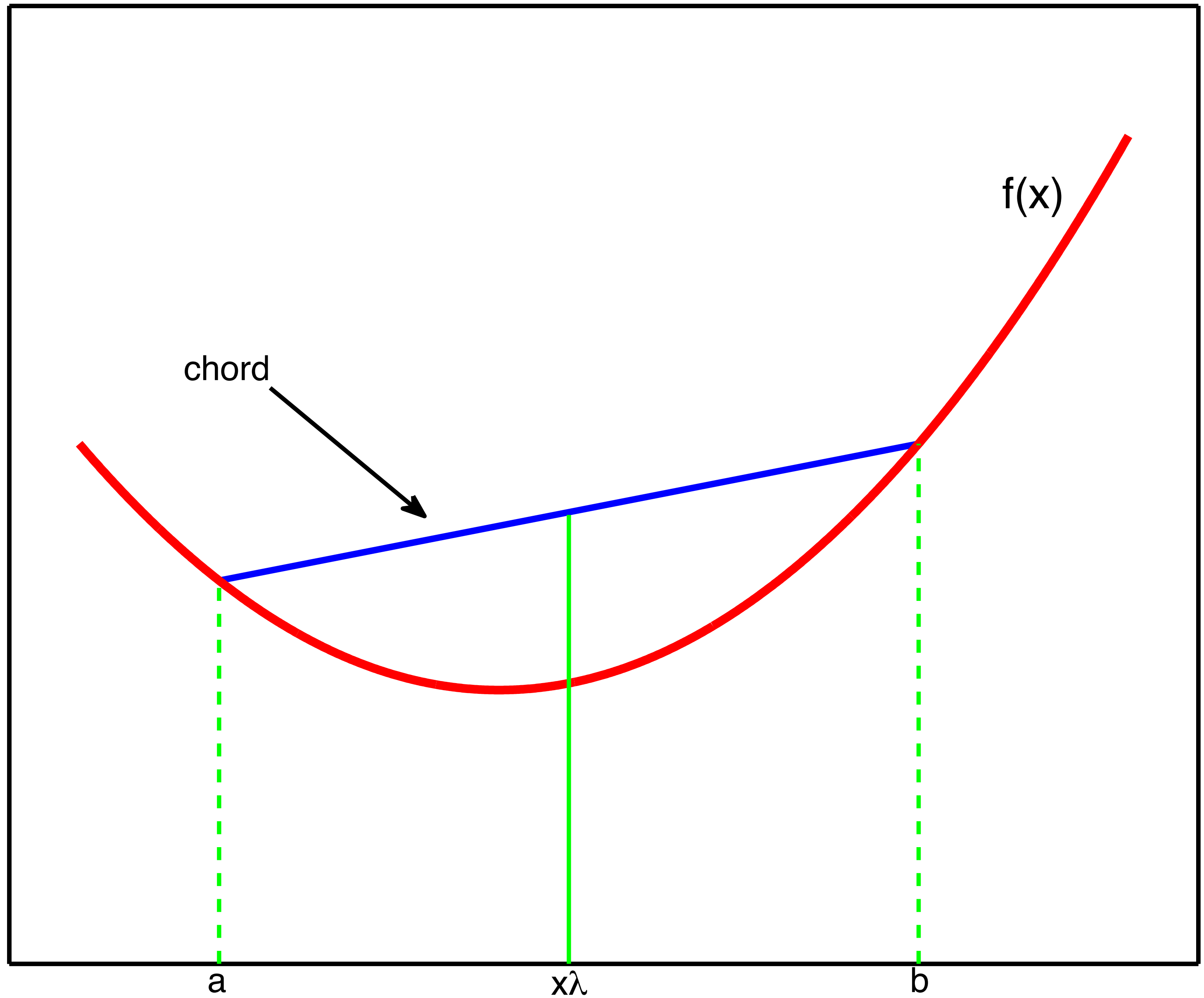
Applied Bayesian Nonparametrics

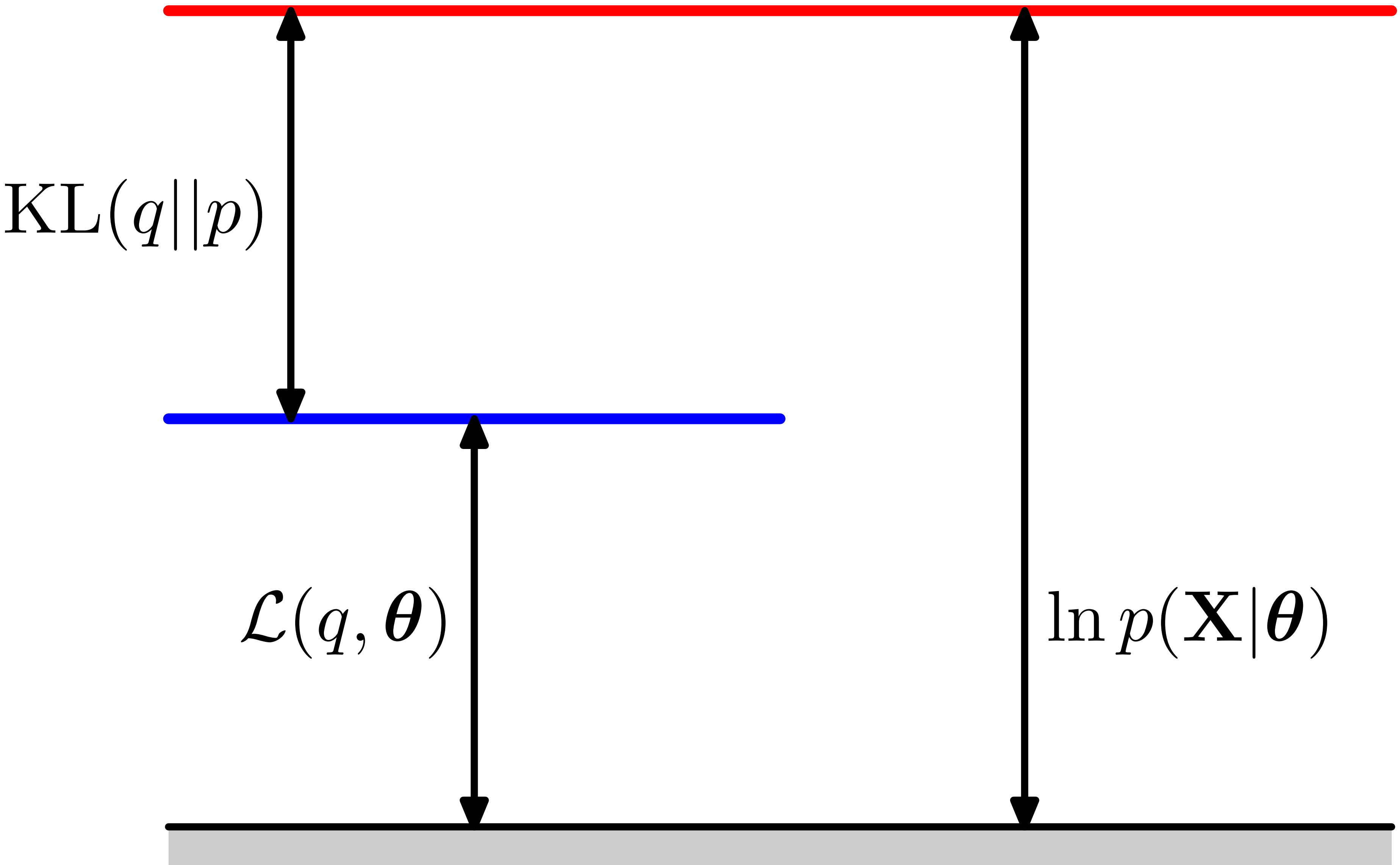
Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

October 11: Variational Methods

Convexity & Jensen's Inequality

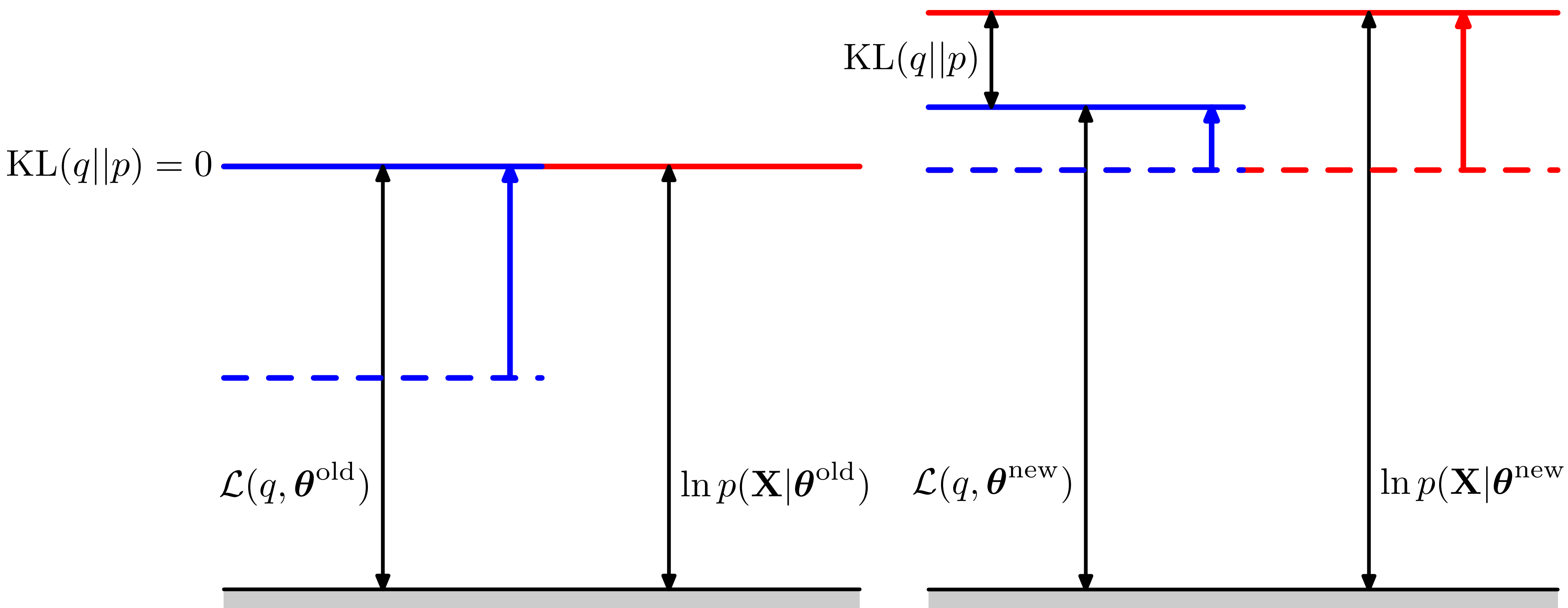


Lower Bounds on Marginal Likelihood



C. Bishop, Pattern Recognition & Machine Learning

Expectation Maximization Algorithm

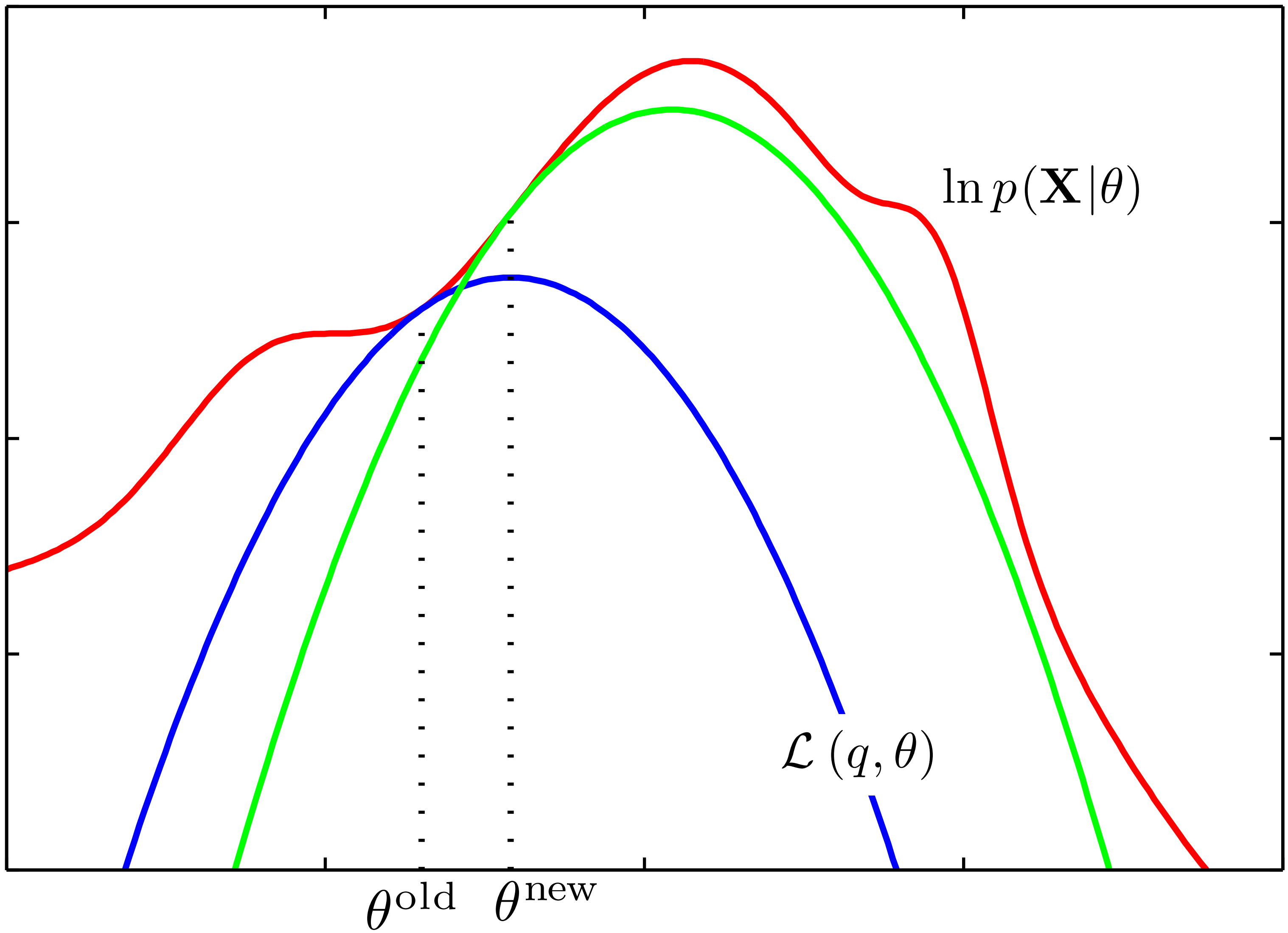


E Step: *Optimize distribution on hidden variables given parameters*

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

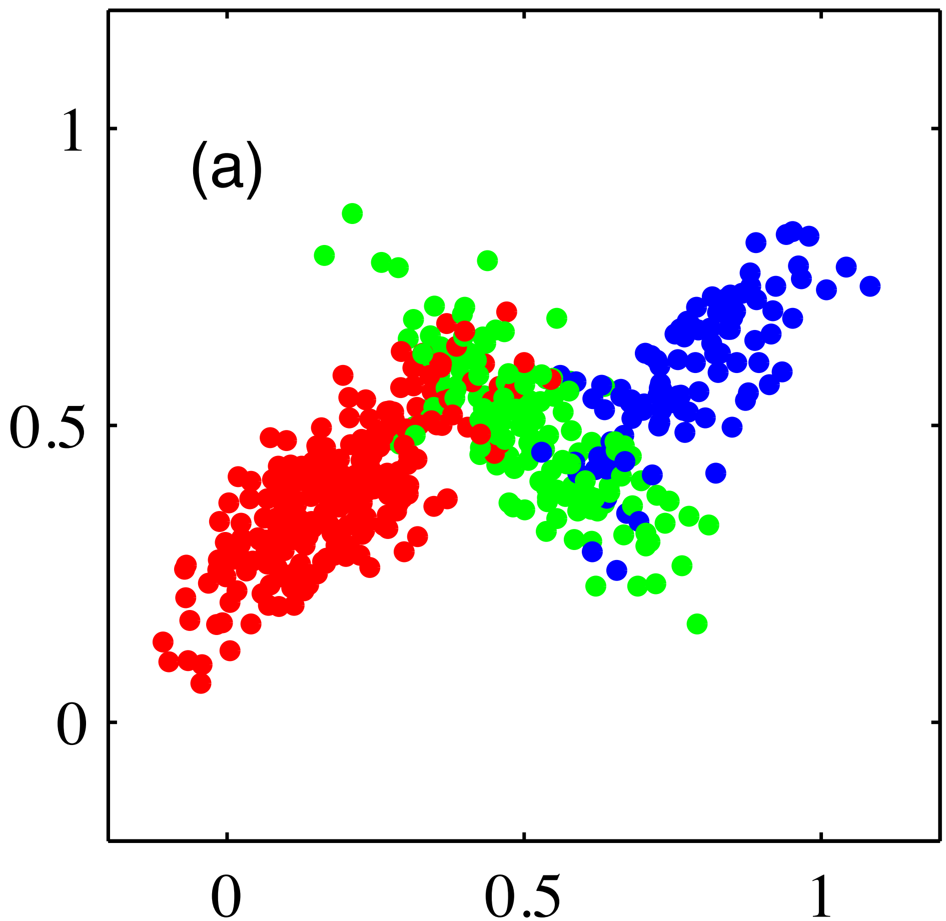
C. Bishop, Pattern Recognition & Machine Learning

EM: A Sequence of Lower Bounds

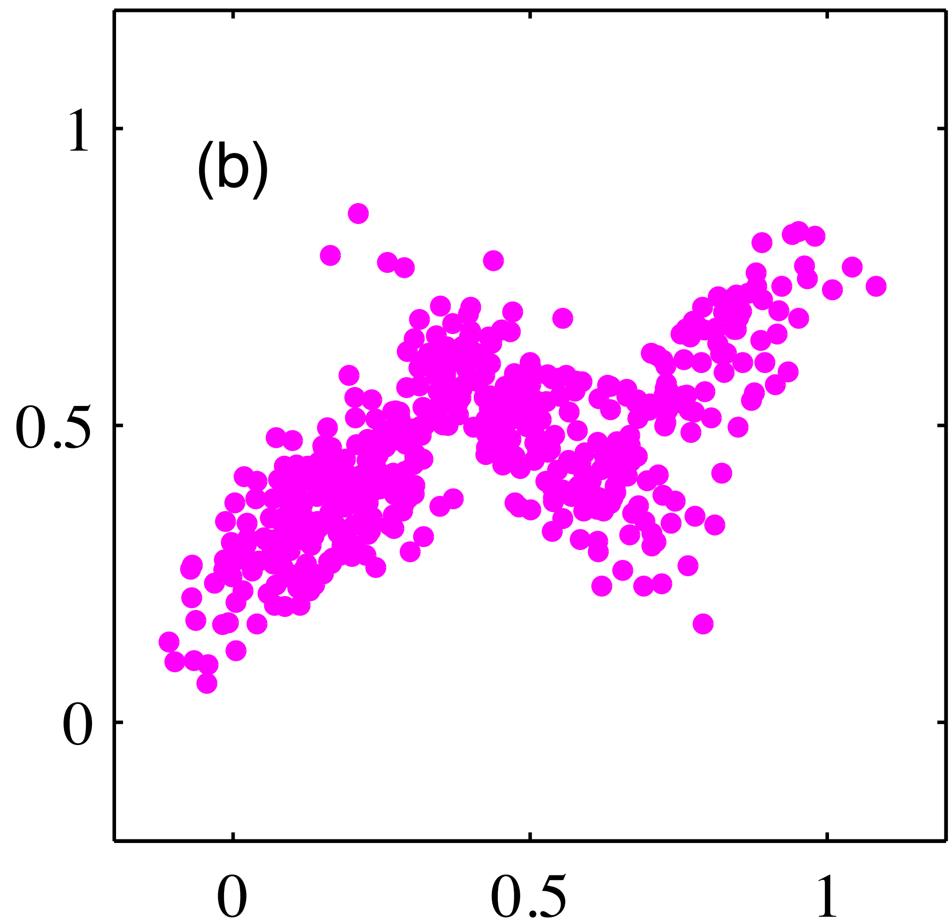


C. Bishop, Pattern Recognition & Machine Learning

Fitting Gaussian Mixtures



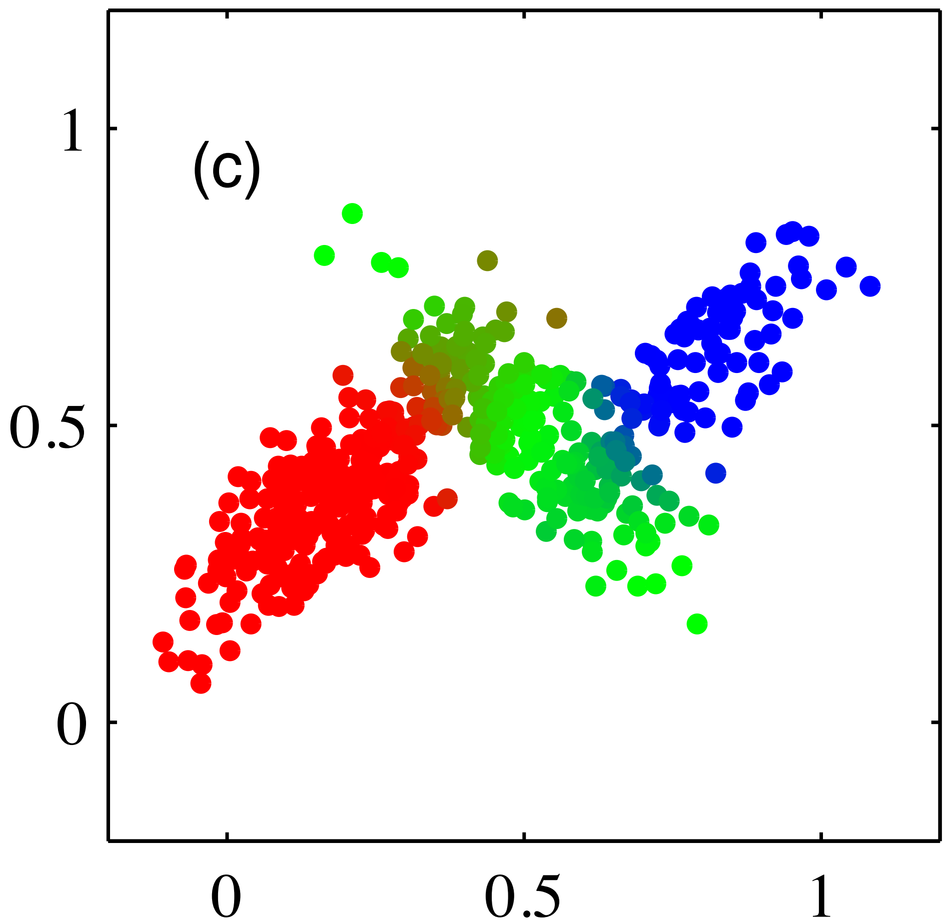
Complete Data Labeled by True Cluster Assignments



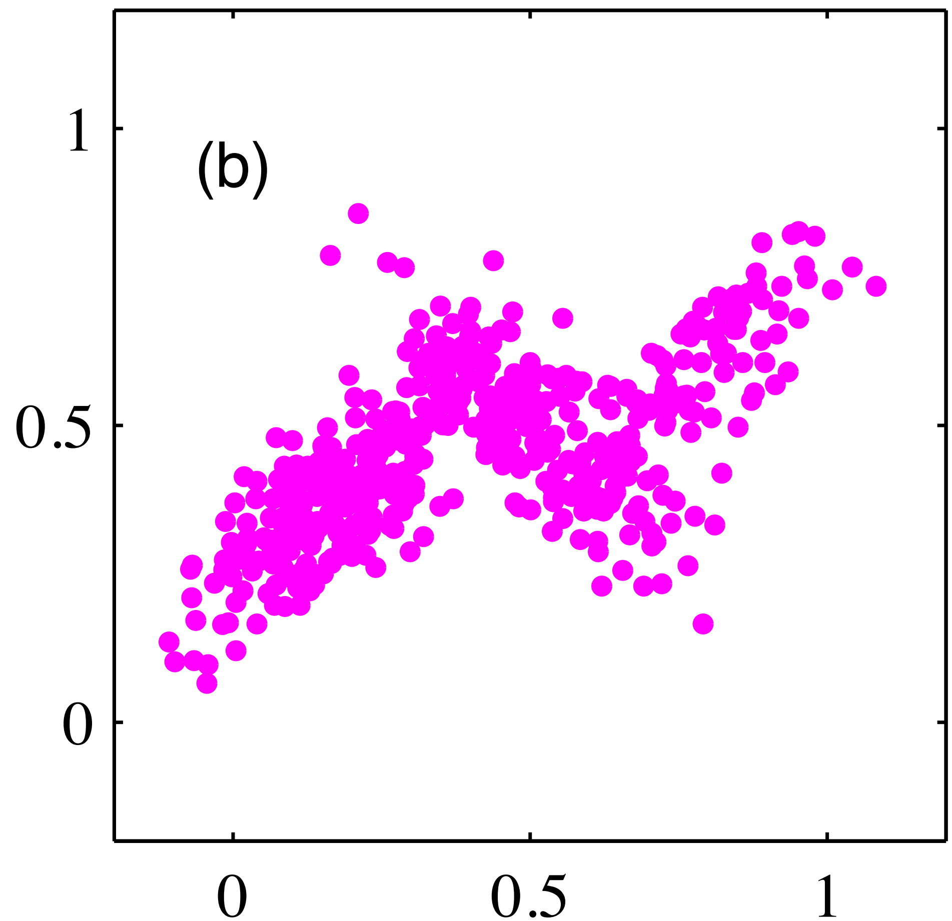
Incomplete Data: Points to be Clustered

C. Bishop, Pattern Recognition & Machine Learning

Posterior Assignment Probabilities



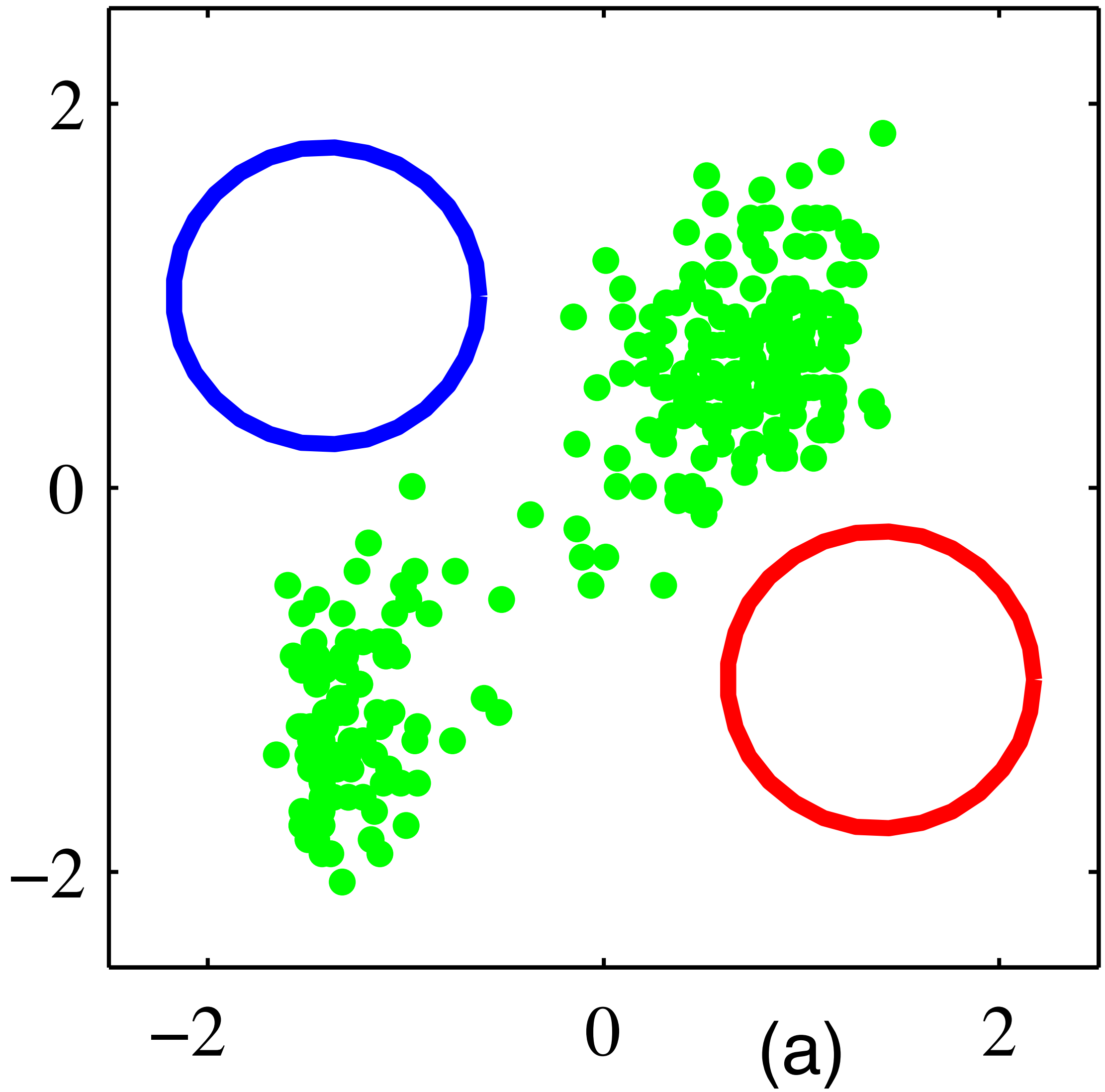
Posterior Probabilities of Assignment to Each Cluster



Incomplete Data: Points to be Clustered

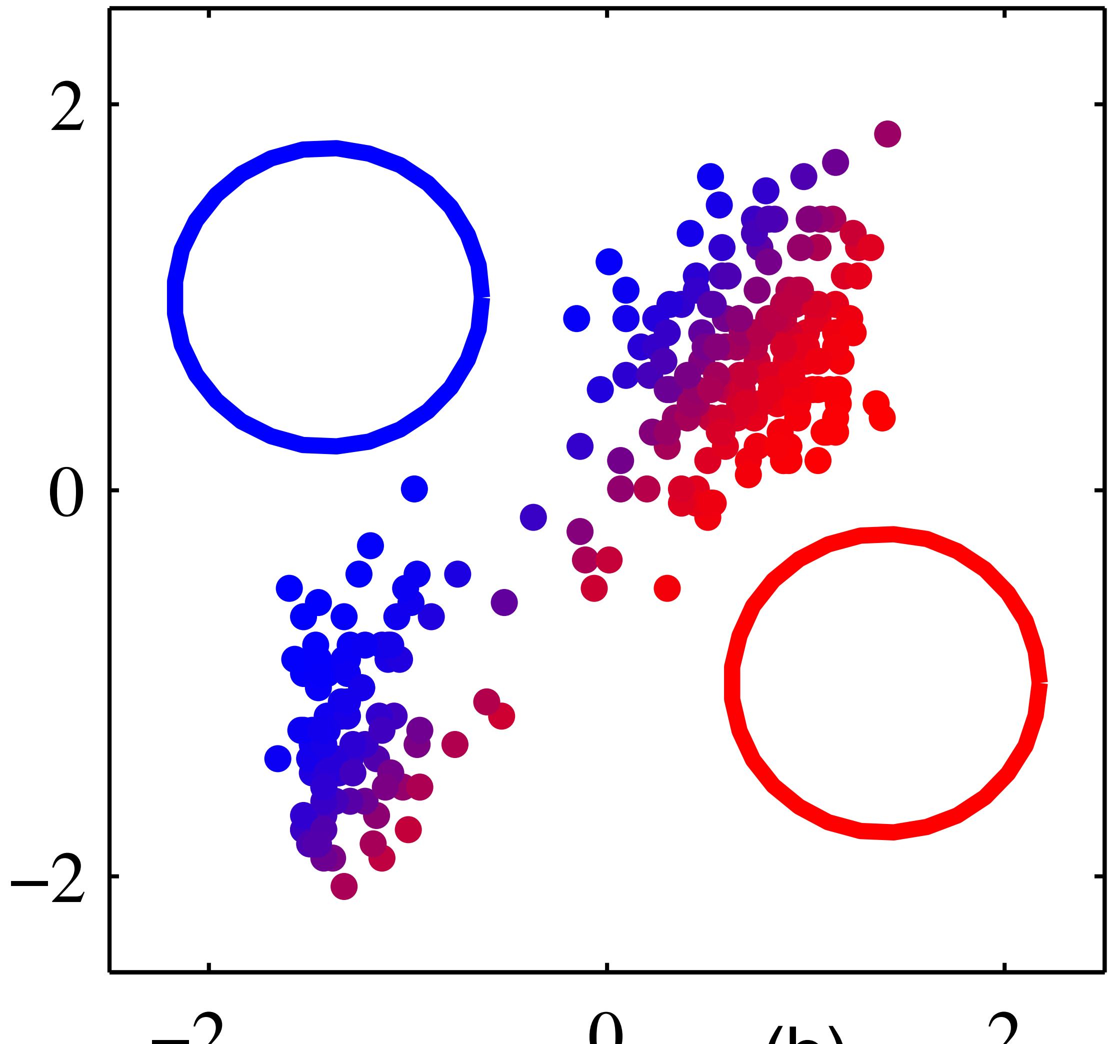
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



C. Bishop, Pattern Recognition & Machine Learning

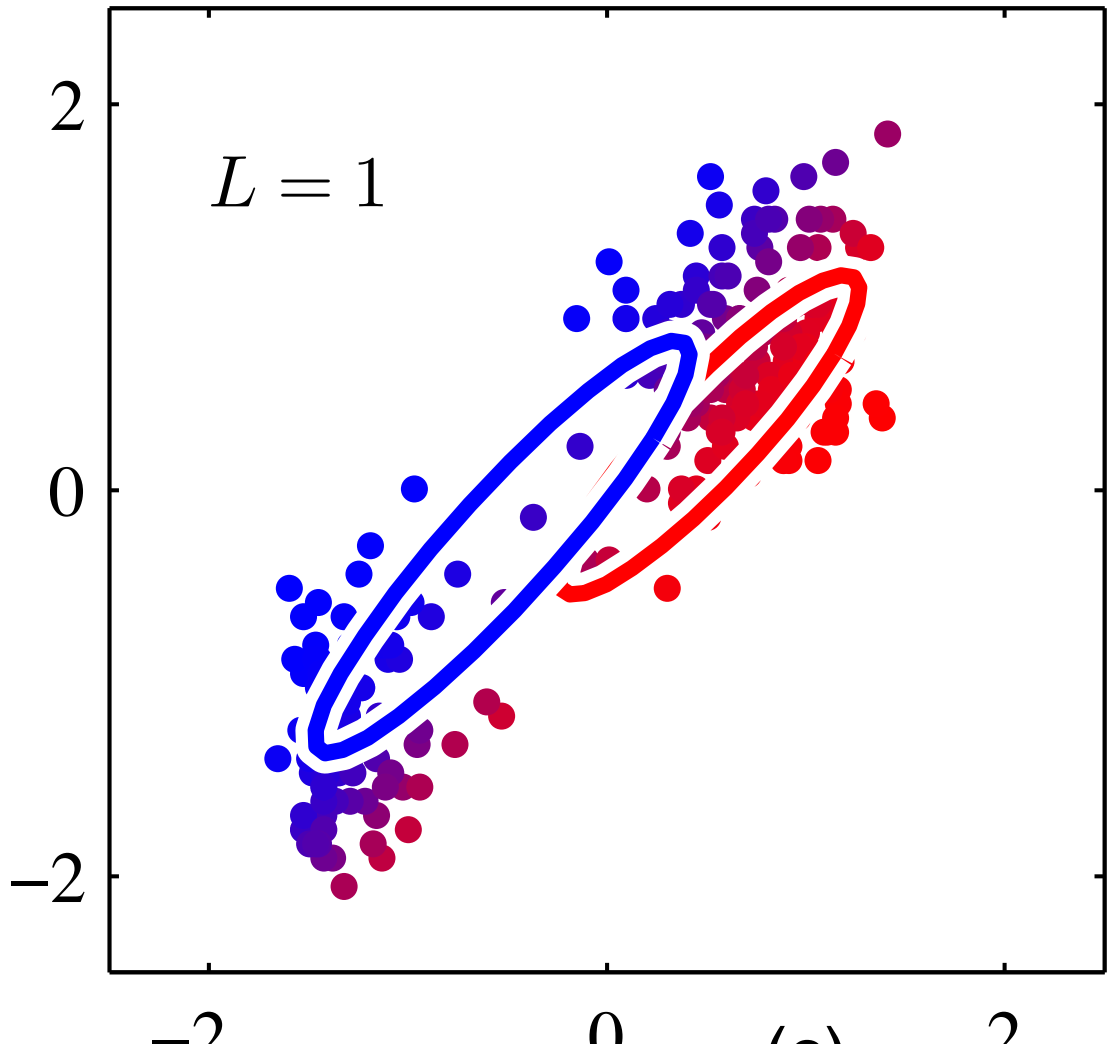
EM Algorithm



(b)

C. Bishop, Pattern Recognition & Machine Learning

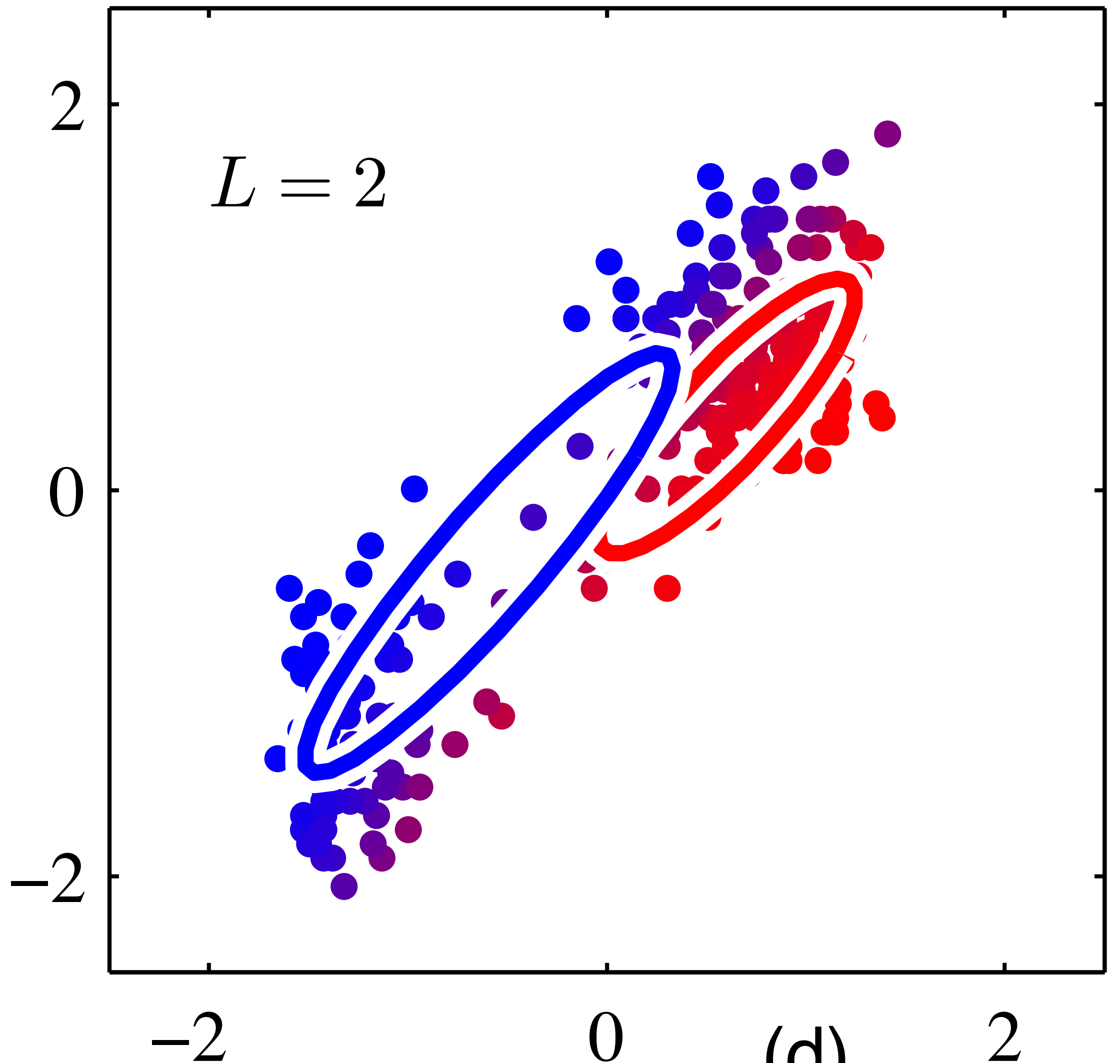
EM Algorithm



(c)

C. Bishop, Pattern Recognition & Machine Learning

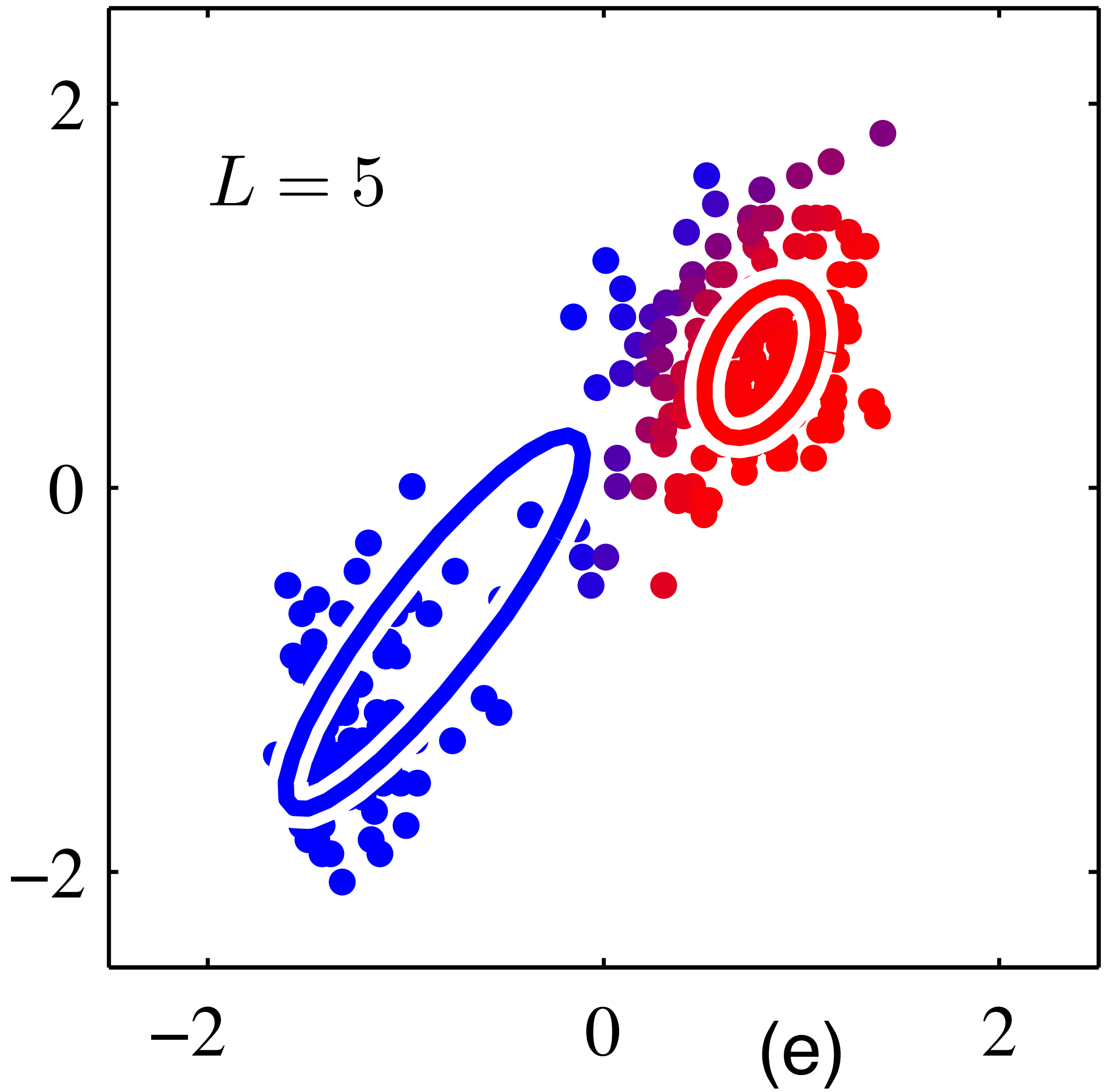
EM Algorithm



(d)

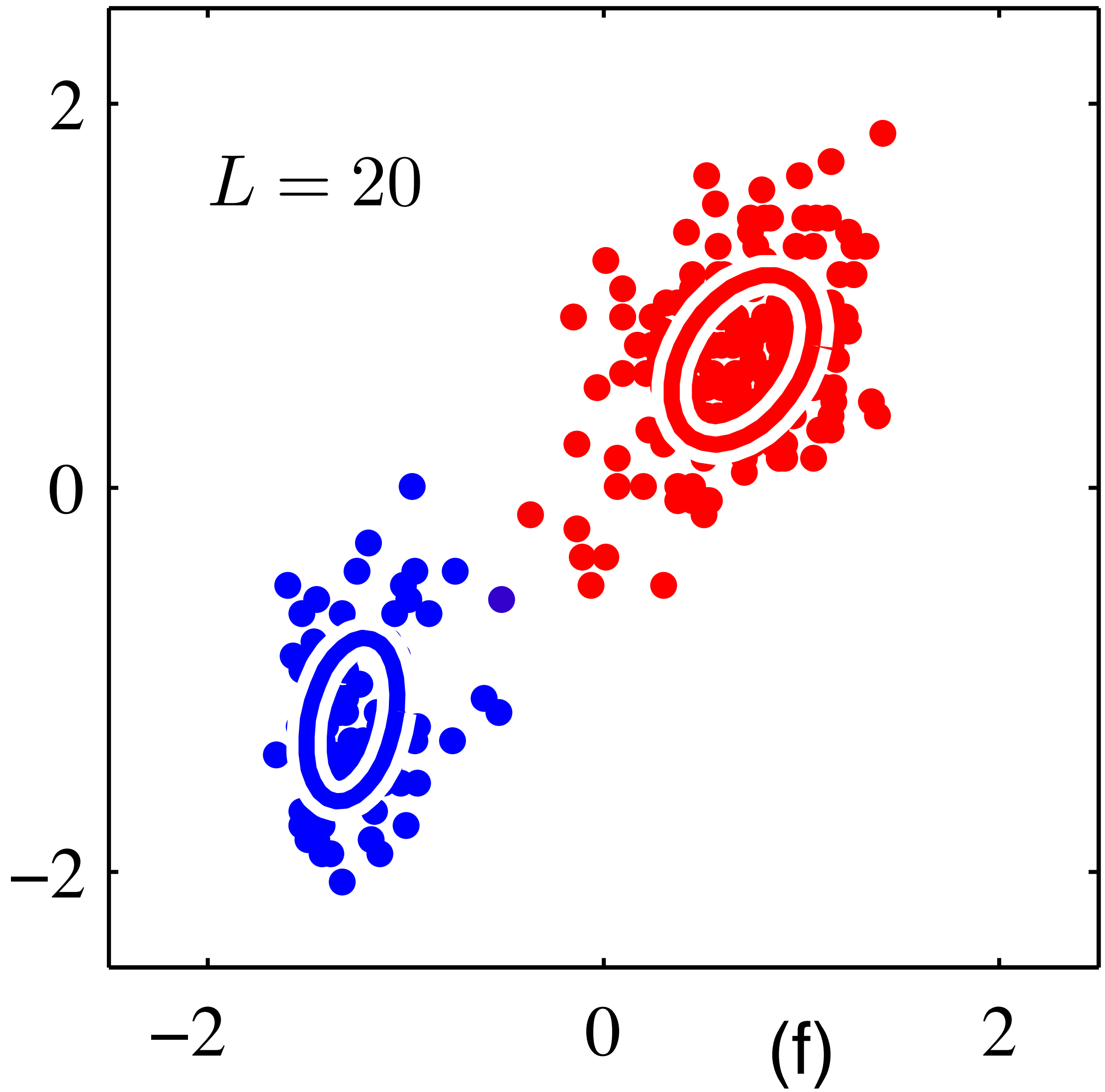
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



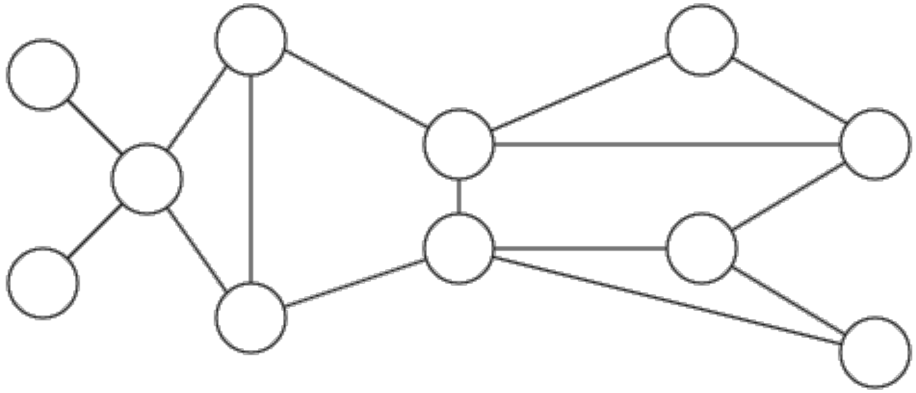
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



C. Bishop, Pattern Recognition & Machine Learning

Pairwise Markov Random Fields



$$p(x \mid y) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s, y)$$

$\mathcal{V}$ ⟶ set of N nodes $\{1, 2, \ldots, N\}$

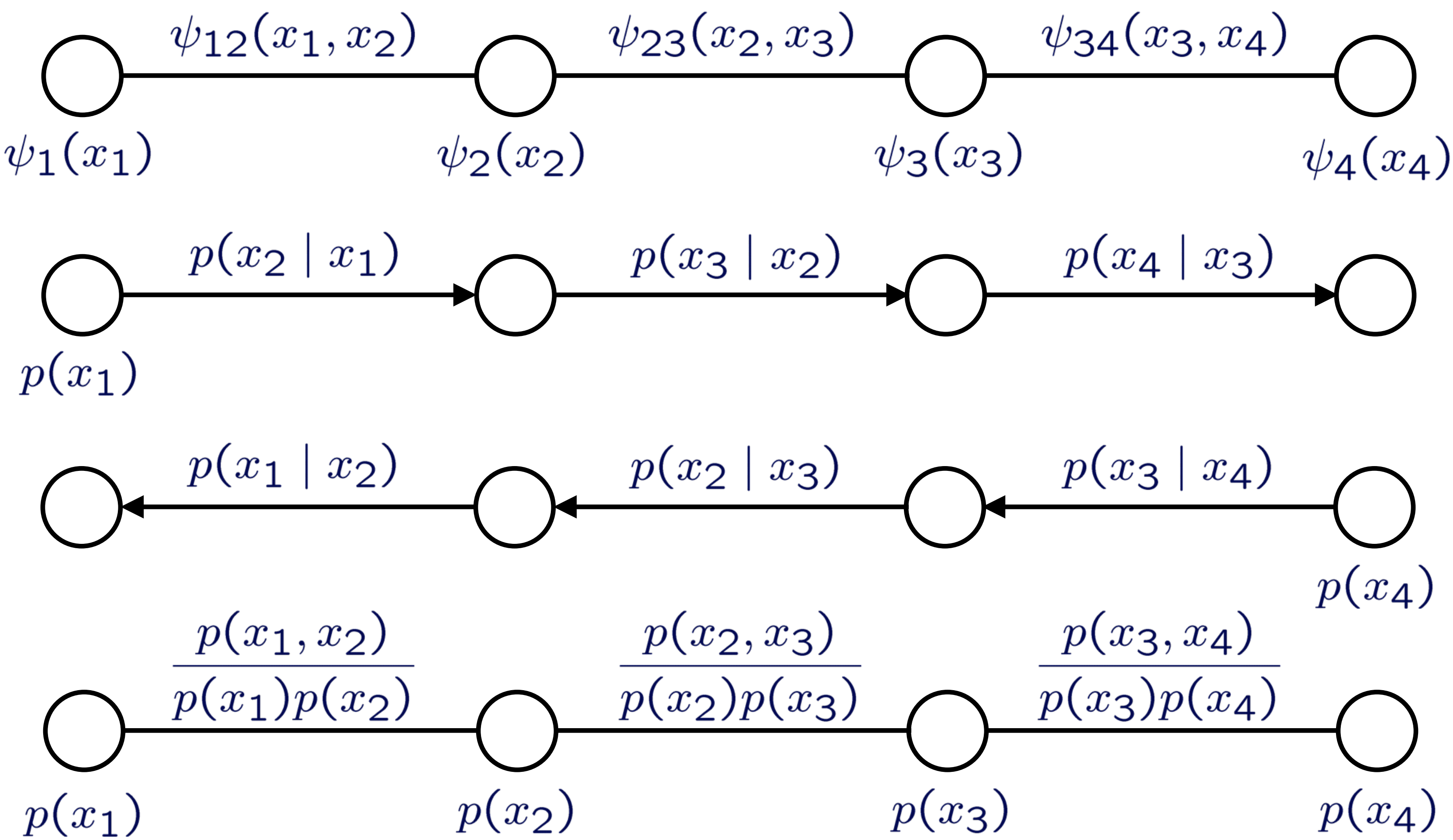
$\mathcal{E}$ ⟶ set of edges (s, t) connecting nodes $s, t \in \mathcal{V}$

Z ⟶ normalization constant (partition function)

- Product of arbitrary positive *clique potential* functions
- Guaranteed Markov with respect to corresponding graph

Markov Chain Factorizations

$$p(x \mid y) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s, y)$$



Energy Functions

$$p(x \mid y) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s, y)$$

$$= \frac{1}{Z} \exp\left\{ -\sum_{(s,t)\in\mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s\in\mathcal{V}} \phi_s(x_s, y) \right\}$$

$$= \frac{1}{Z} \exp\{-E(x)\}$$

$$\phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t) \qquad \phi_s(x_s) = -\log \psi_s(x_s)$$

Interpretation and terminology from statistical physics

Approximate Inference Framework

$$p(x \mid y) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s, y)$$

- Choose a family of approximating distributions which is tractable. The simplest example:

$$q(x) = \prod_{s \in \mathcal{V}} q_s(x_s)$$

- Define a distance to measure the quality of different approximations. Two possibilities:

$$D(p \,||\, q) = \sum_x p(x \mid y) \log \frac{p(x \mid y)}{q(x)}$$

$$D(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x \mid y)}$$

- Find the approximation minimizing this distance

Fully Factored Approximations

$$p(x \mid y) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s, y)$$

$$q(x) = \prod_{s\in\mathcal{V}} q_s(x_s)$$

$$D(p \mid\mid q) = \sum_x p(x \mid y) \log \frac{p(x \mid y)}{q(x)}$$

$$= \left[\sum_{s\in\mathcal{V}} H_s(p_s) - H(p) \right] + \sum_{s\in\mathcal{V}} D(p_s \mid\mid q_s)$$

Marginal Entropies

Joint Entropy

- Trivially minimized by setting $q_s(x_s) = p_s(x_s \mid y)$
- Doesn't provide a computational method…

Variational Approximations

$$D(q(x) \,||\, p(x \mid y)) = \sum_x q(x) \log \frac{q(x)}{p(x \mid y)}$$

$$\begin{aligned}
\log p(y) &= \log \sum_x p(x, y) \\
&= \log \sum_x q(x) \frac{p(x, y)}{q(x)} && \textit{(Multiply by one)} \\
&\geq \underbrace{\sum_x q(x) \log \frac{p(x, y)}{q(x)}}_{} && \textit{(Jensen's inequality)} \\
&= -D(q(x) \,||\, p(x \mid y)) + \log p(y)
\end{aligned}$$

- Minimizing KL divergence maximizes a lower bound on the data likelihood

Free Energies

$$p(x \mid y) = \frac{1}{Z} \exp\{-E(x)\}$$

$$D(q \,||\, p) = \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x \mid y)$$

$$= \underbrace{\underbrace{-H(q)}_{\text{Negative Entropy}} + \underbrace{\sum_x q(x) E(x)}_{\text{Average Energy}}}_{\text{Gibbs Free Energy}} + \underbrace{\log Z}_{\text{Normalization}}$$

Negative Entropy | *Average Energy* | *Normalizat ion*

Gibbs Free Energy

- Free energies equivalent to KL divergence, up to a normalization constant

Mean Field Free Energy

$$p(x \mid y) = \frac{1}{Z} \exp \left\{ - \sum_{(s,t) \in \mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s \in \mathcal{V}} \phi_s(x_s, y) \right\}$$

$$q(x) = \prod_{s \in \mathcal{V}} q_s(x_s)$$

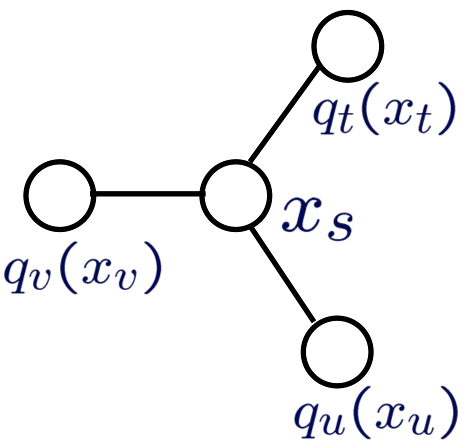
$$\begin{aligned} D(q \,||\, p) &= -H(q) + \sum_x q(x) E(x) + \log Z \\ &= -\sum_{s \in \mathcal{V}} H_s(q_s) + \sum_{(s,t) \in \mathcal{E}} q_s(x_s) q_t(x_t) \phi_{st}(x_s, x_t) \\ &\quad \cdots + \sum_{s \in \mathcal{V}} q_s(x_s) \phi_s(x_s) + \log Z \end{aligned}$$

Mean Field Equations

$$D(q \,||\, p) = -\sum_{s\in\mathcal{V}} H_s(q_s) + \sum_{(s,t)\in\mathcal{E}} q_s(x_s)q_t(x_t)\phi_{st}(x_s,x_t)$$
$$\cdots + \sum_{s\in\mathcal{V}} q_s(x_s)\phi_s(x_s) + \log Z$$

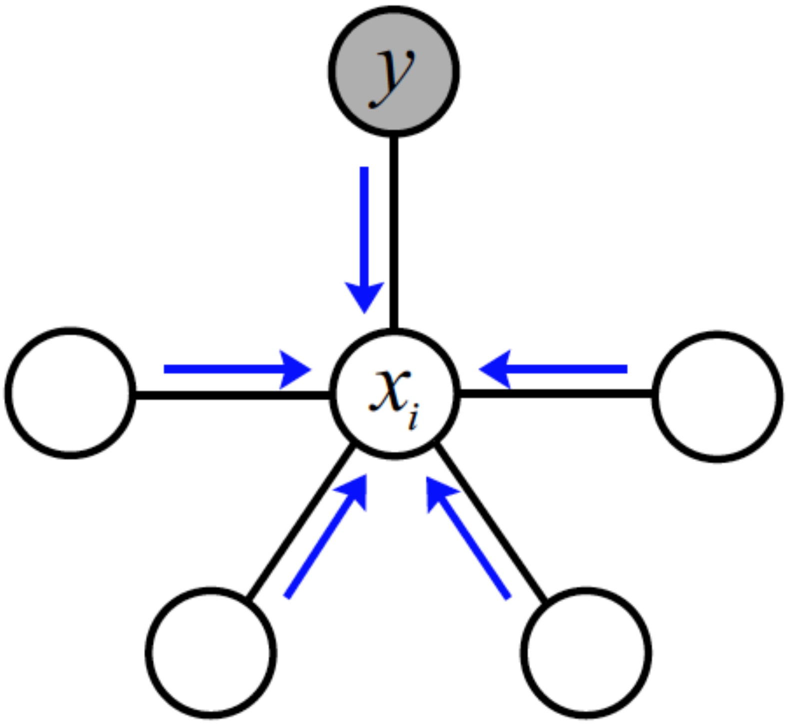
- Add Lagrange multipliers to enforce $\sum_{x_s} q_s(x_s) = 1$
- Taking derivatives and simplifying, we find a set of fixed point equations:

$$q_s(x_s) = \alpha\psi_s(x_s) \prod_{t\in\Gamma(s)} \prod_{x_t} \psi_{st}(x_s,x_t)^{q_t(x_t)}$$



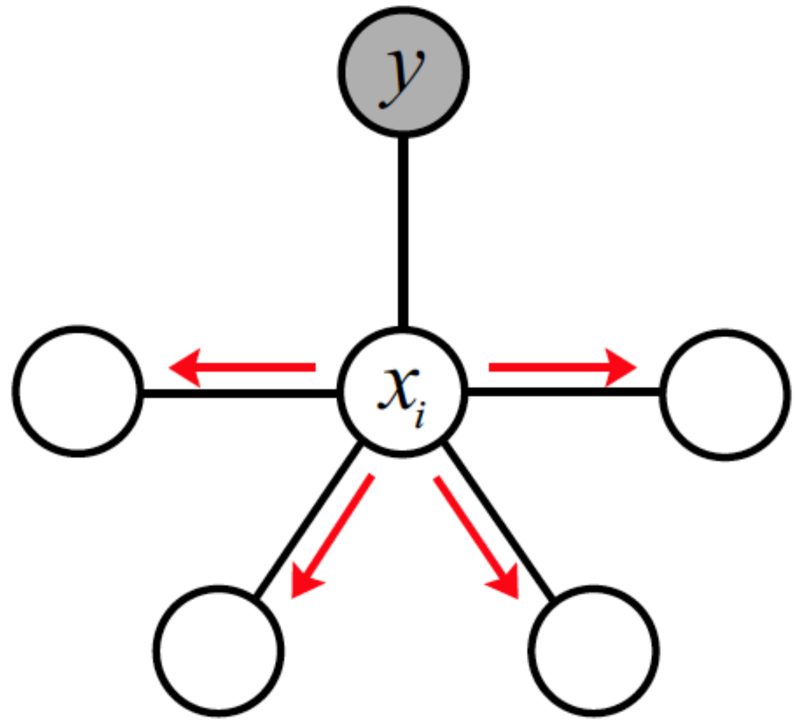
- Updating one marginal at a time gives convergent coordinate descent

Mean Field Message Passing



$$q_i(x_i) \propto \psi_i(x_i, y) \prod_{j \in \Gamma(i)} m_{ji}(x_i)$$

Want products of messages to be simple



$$m_{ij}(x_j) \propto \exp\left\{ - \int_{\mathcal{X}_i} \phi_{ji}(x_j, x_i)\, q_i(x_i)\, dx_i \right\}$$

Want expectations of log potential functions to be simple

Exponential Families

- Natural or canonical parameters determine log-linear combination of sufficient statistics:

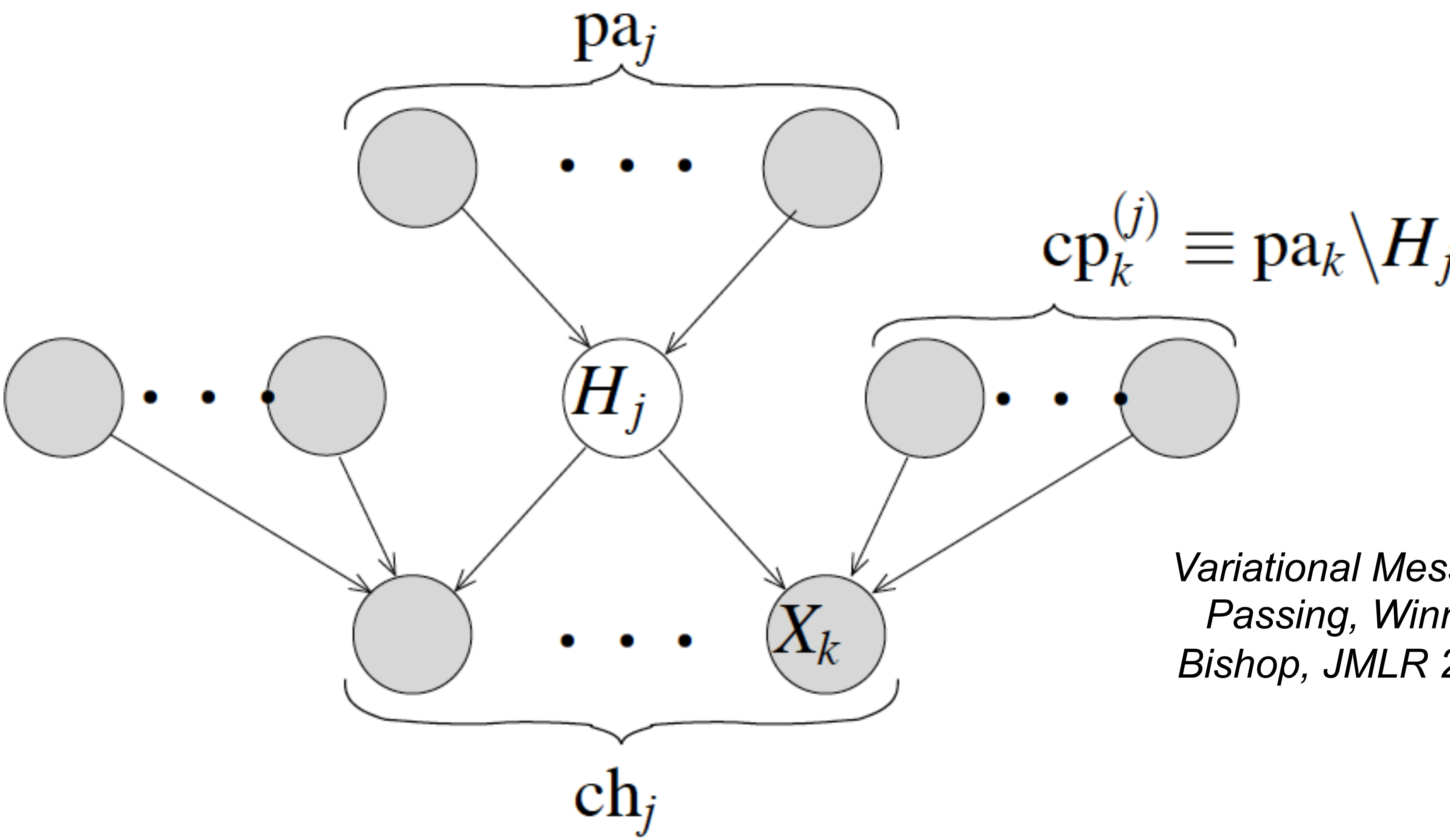
$$p(x \mid \theta) = \nu(x) \exp\left\{ \sum_{a \in \mathcal{A}} \theta_a \phi_a(x) - \Phi(\theta) \right\}$$

- Log partition function normalizes to produce valid probability distribution:

$$\Phi(\theta) = \log \int_{\mathcal{X}} \nu(x) \exp\left\{ \sum_{a \in \mathcal{A}} \theta_a \phi_a(x) \right\} dx$$

$$\Theta \triangleq \left\{ \theta \in \mathbb{R}^{|\mathcal{A}|} \mid \Phi(\theta) < \infty \right\}$$

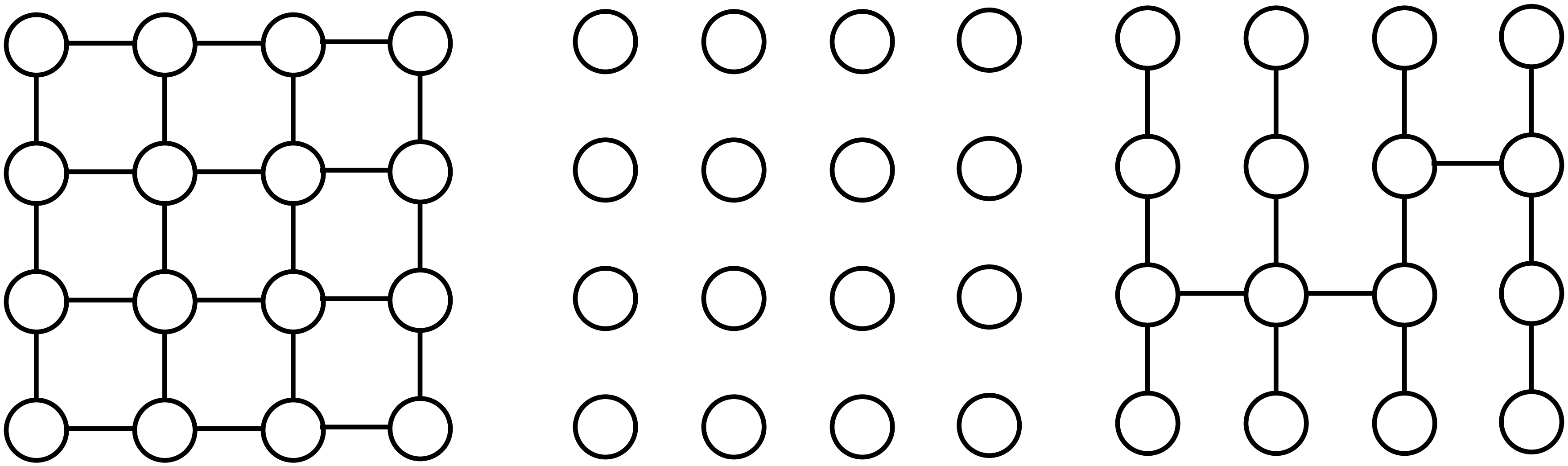
Directed Mean Field



Variational Message Passing, Winn & Bishop, JMLR 2005

- Can derive updates using exponential family form of the conditional distribution of each variable, given its parents
- Can also just take derivatives, collect terms, simplify…

Structured Mean Field



Original Graph **Naïve Mean Field** **Structured Mean Field**

- Any subgraph for which inference is tractable leads to a mean field style approximation for which the update equations are tractable