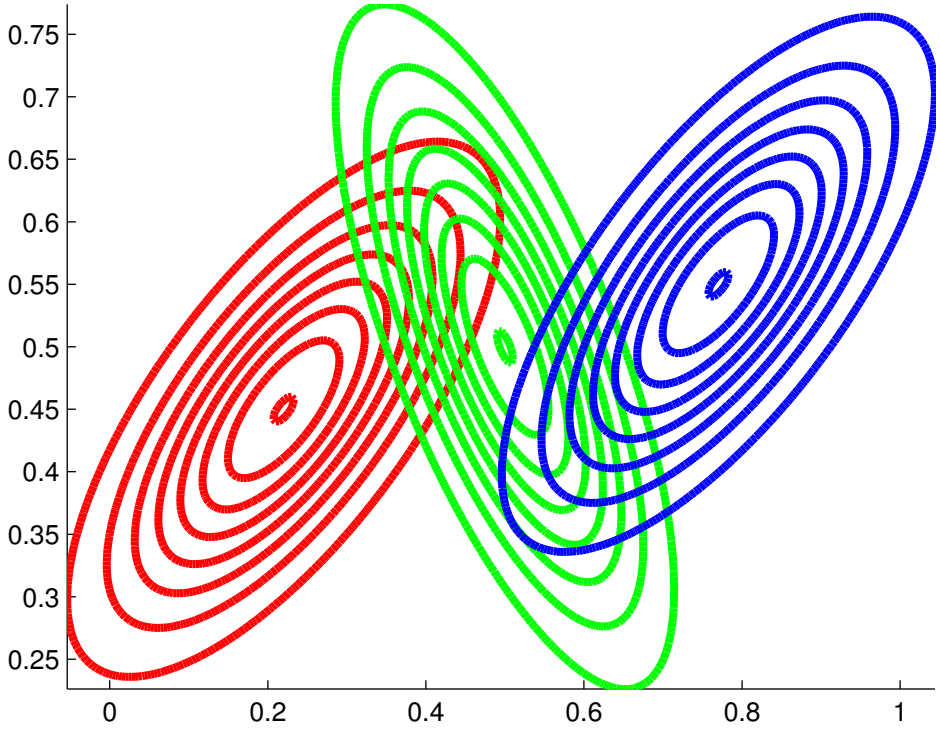
Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

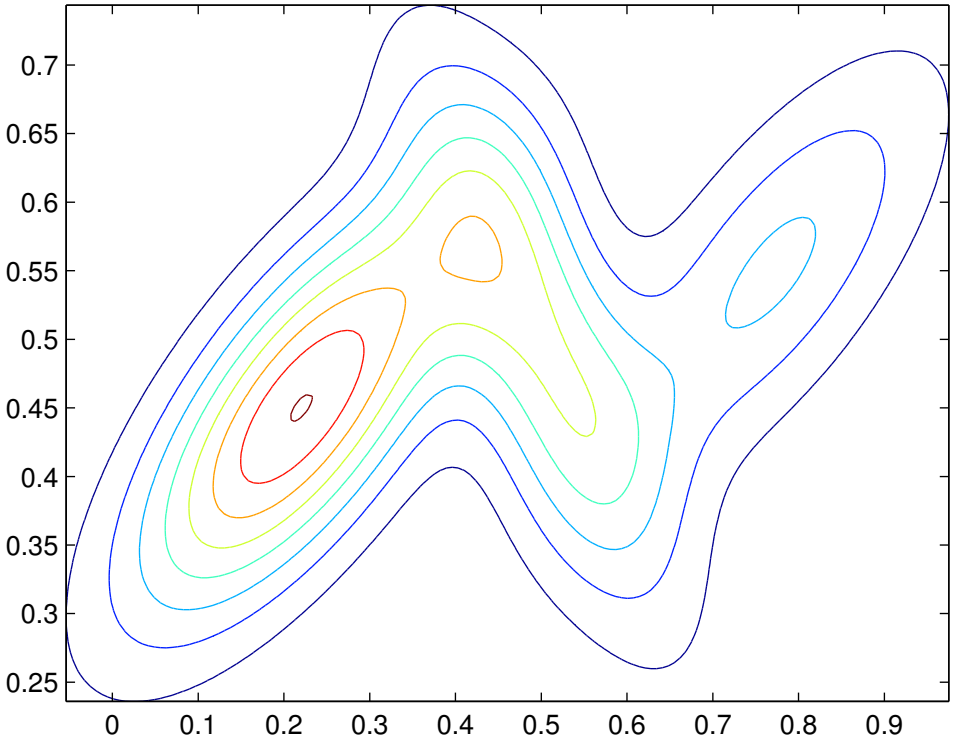
Lecture 20:
Expectation Maximization Algorithm
EM for Mixture Models

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Gaussian Mixture Models



Mixture of 3 Gaussian Distributions in 2D



Contour Plot of Joint Density, Marginalizing Cluster Assignments

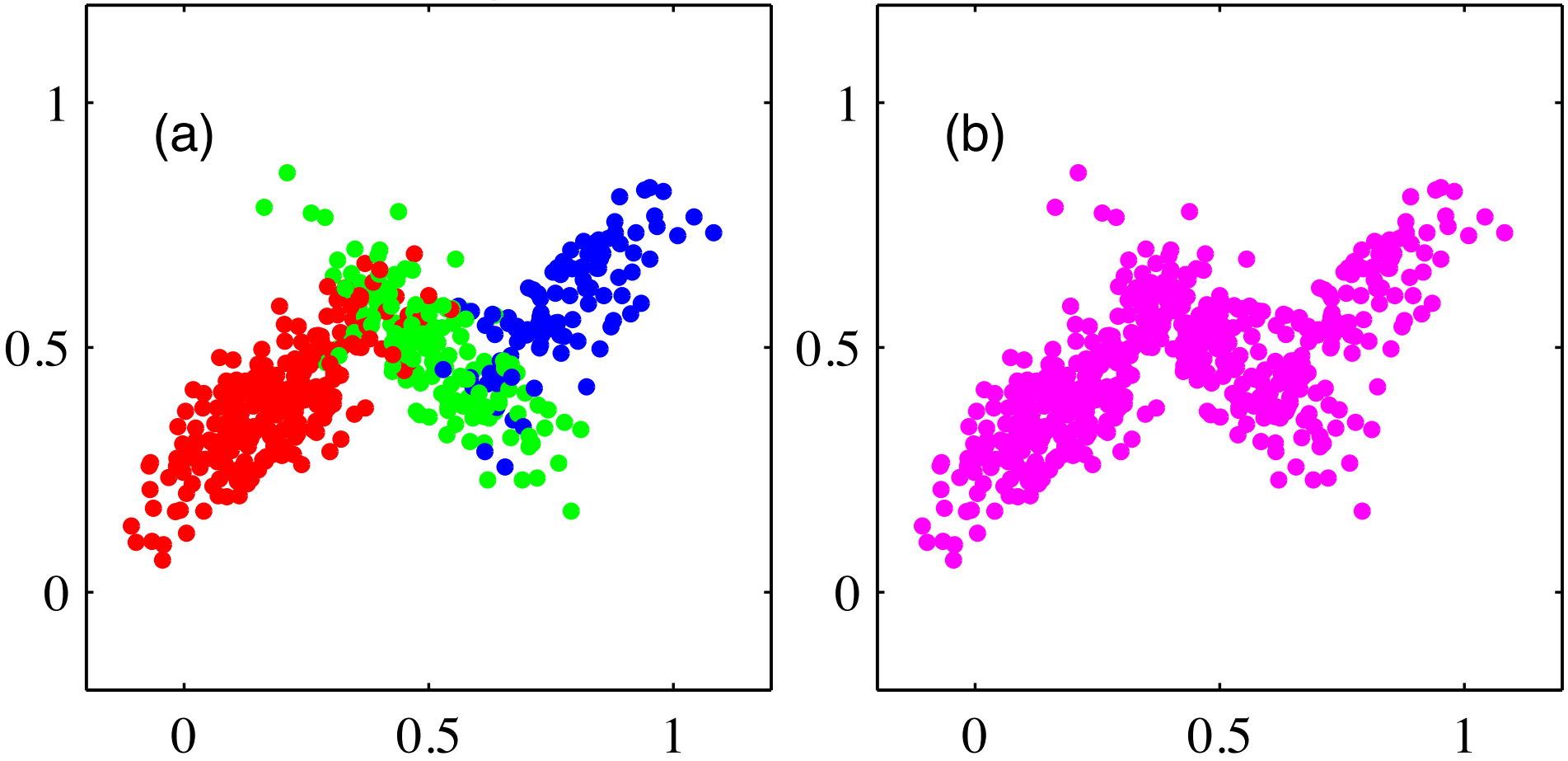
Gaussian Mixture Models

- Observed feature vectors: $x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N$
- Hidden cluster labels: $z_i \in \{1, 2, \ldots, K\}, \quad i = 1, 2, \ldots, N$
- Hidden mixture means: $\mu_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, K$
- Hidden mixture covariances: $\Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \ldots, K$
- Hidden mixture probabilities: $\pi_k, \quad \sum_{k=1}^{K} \pi_k = 1$
- Gaussian mixture marginal likelihood:

$$p(x_i \mid \pi, \mu, \Sigma) = \sum_{z_i=1}^{K} \pi_{z_i} \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

$$p(x_i \mid z_i, \pi, \mu, \Sigma) = \mathcal{N}(x_i \mid \mu_{z_i}, \Sigma_{z_i})$$

Clustering with Gaussian Mixtures



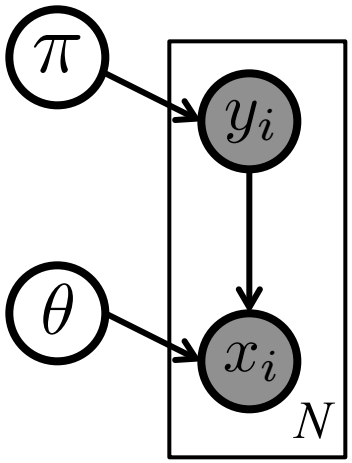
Complete Data Labeled by True Cluster Assignments

Incomplete Data: Points to be Clustered

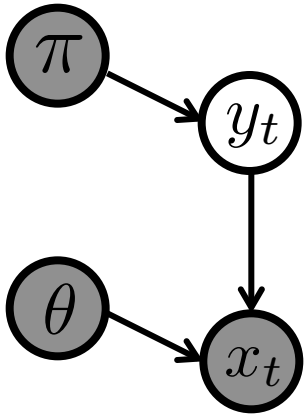
With complete data, learning is Gaussian discriminant analysis.

C. Bishop, Pattern Recognition & Machine Learning

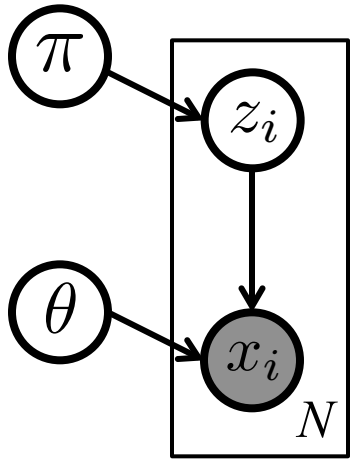
Expectation Maximization (EM)



Supervised Training



Supervised Testing



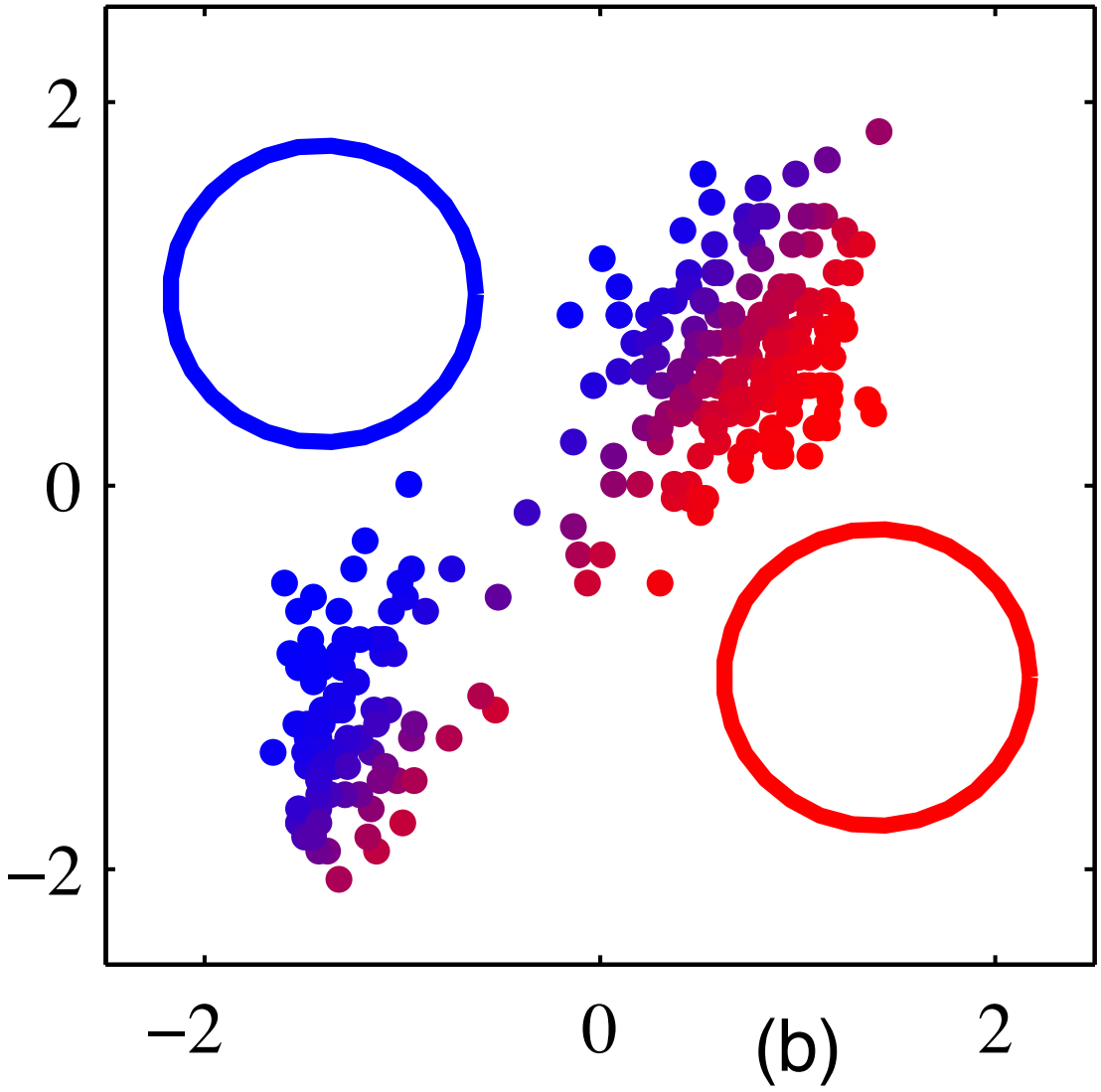
Unsupervised Learning

$\pi, \theta \longrightarrow$ *parameters (define cluster location and shape)*

$z_1, \ldots, z_N \longrightarrow$ *hidden data (group observations into clusters)*

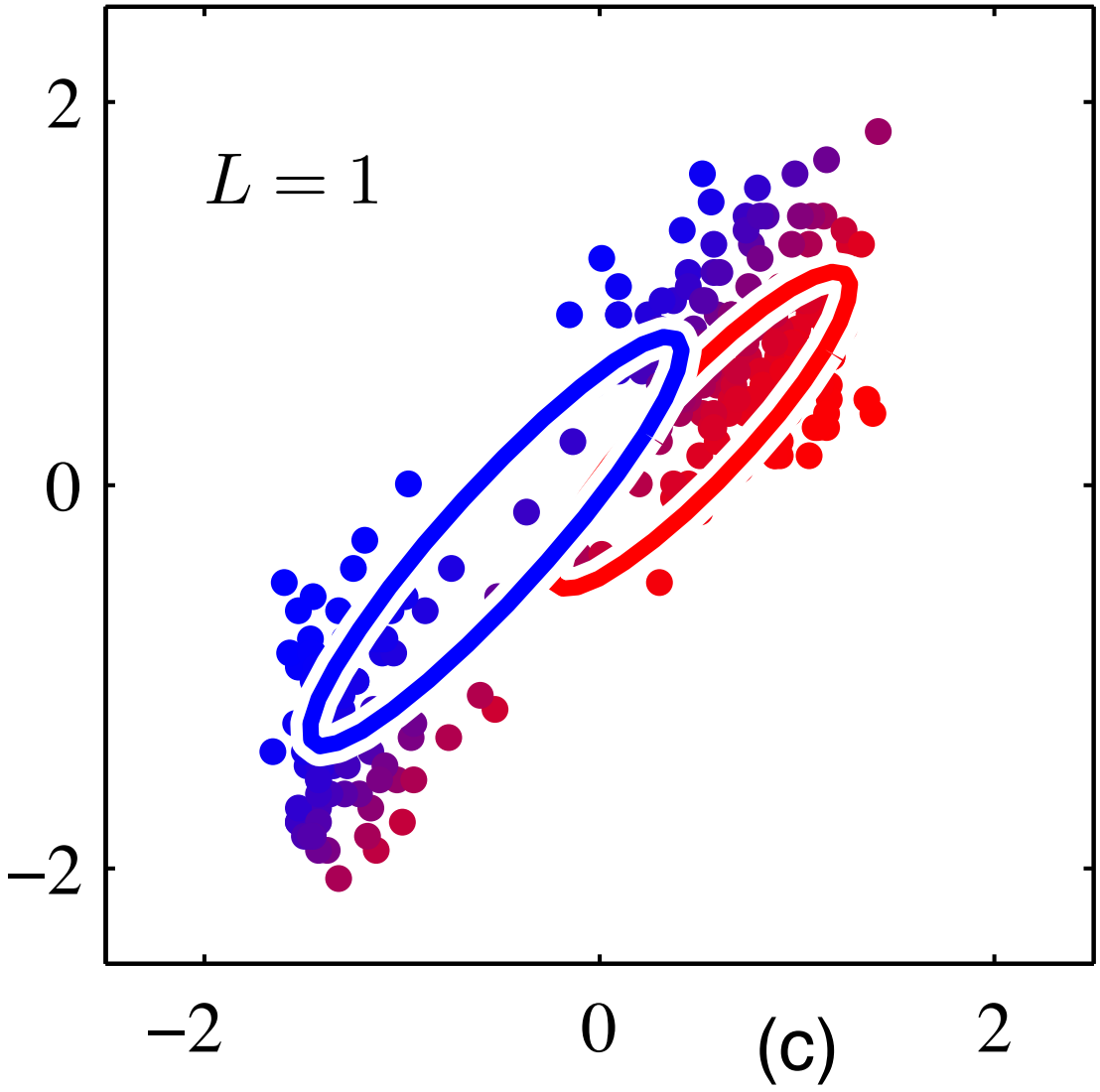
- **Initialization:** Randomly select starting parameters
- **E-Step:** Given parameters, find posterior of hidden data
  - Equivalent to test inference of full posterior distribution
- **M-Step:** Given posterior distributions, find likely parameters
  - Distinct from supervised ML/MAP, but often still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

EM Algorithm



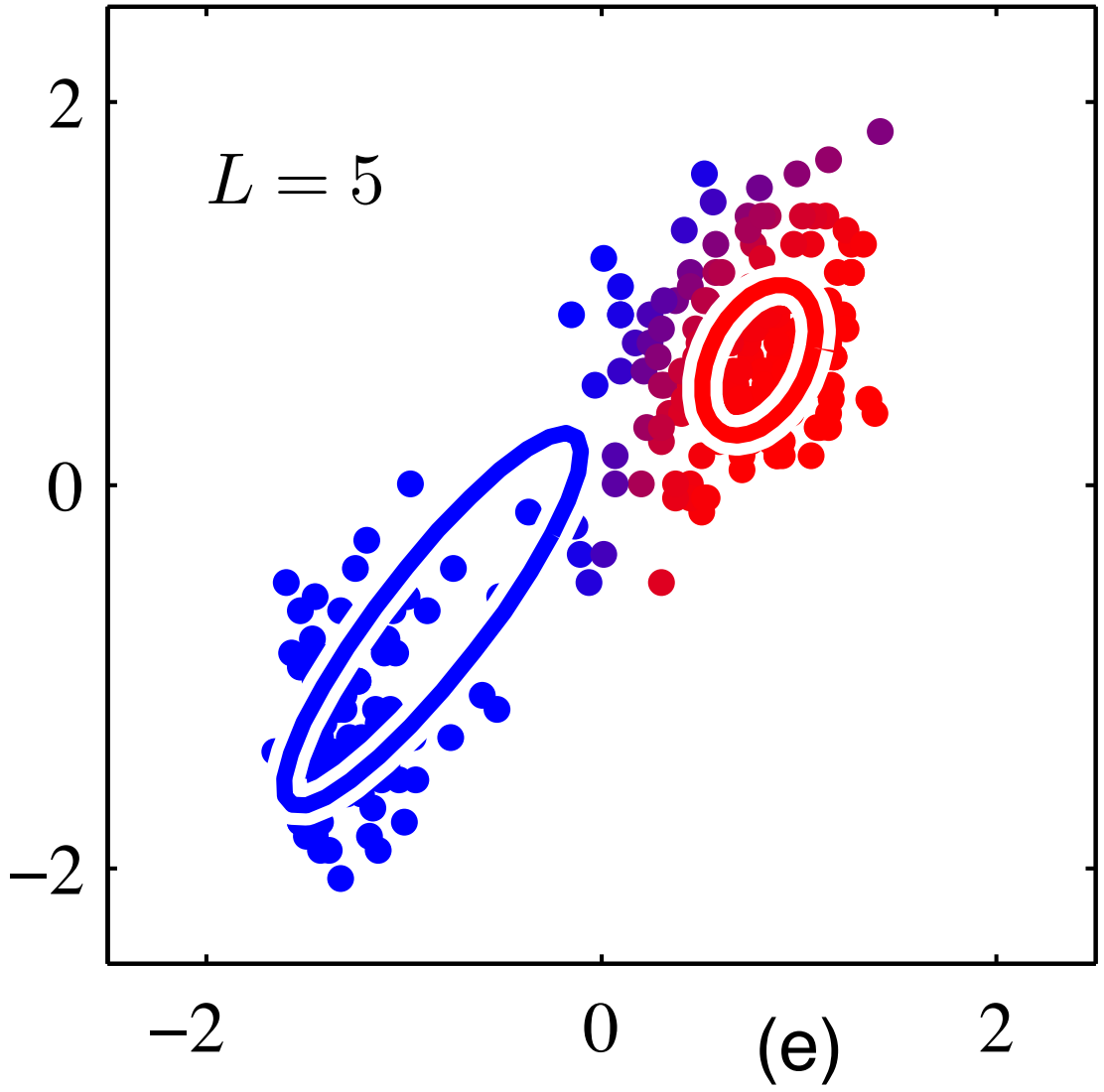
C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm



C. Bishop, Pattern Recognition & Machine Learning

EM Algorithm

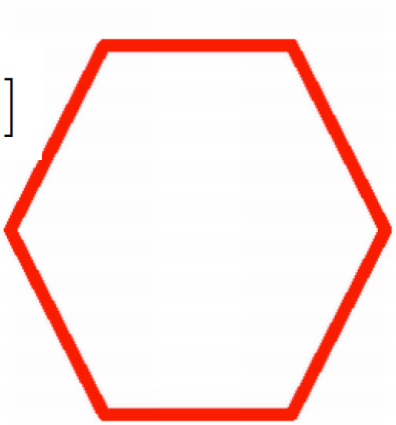


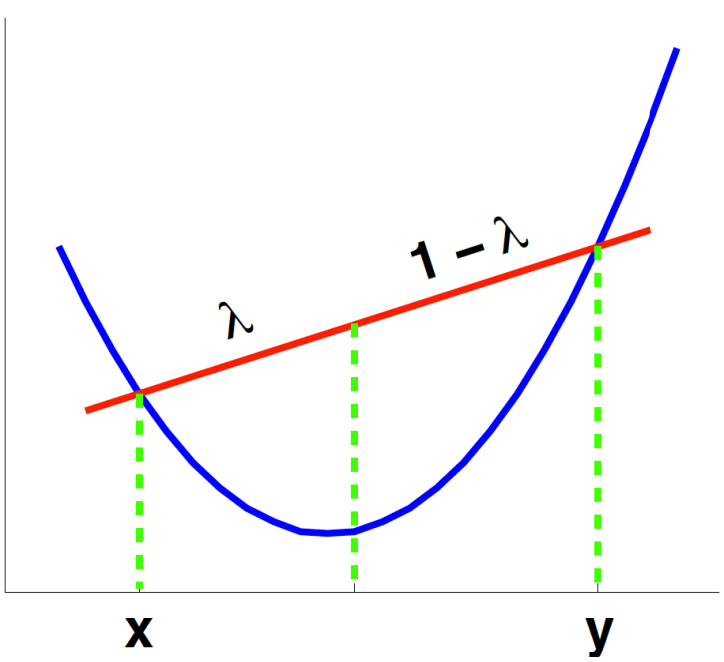
C. Bishop, Pattern Recognition & Machine Learning

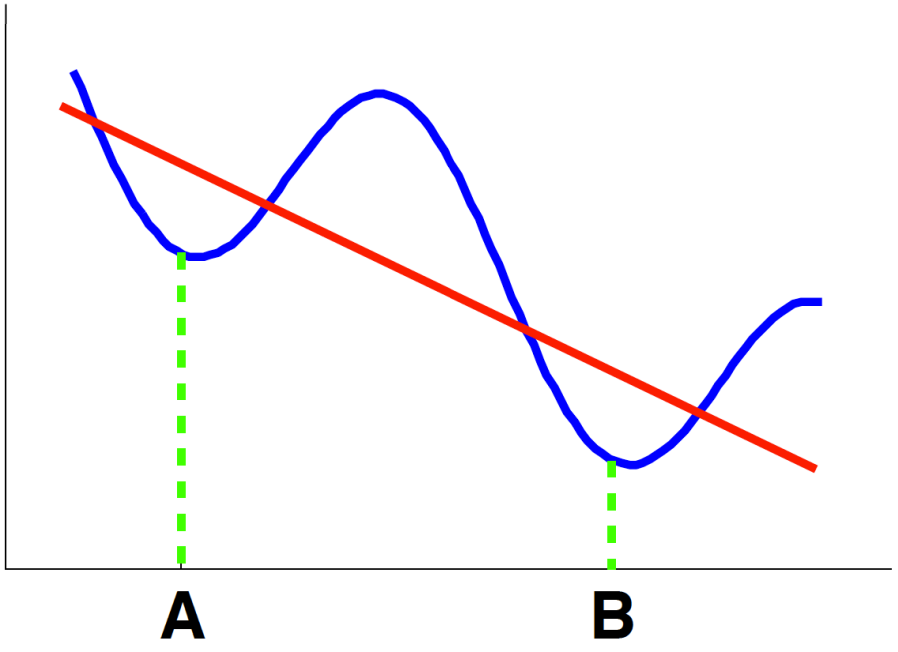
Convexity

$$\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}' \in \mathcal{S}, \quad \forall\, \lambda \in [0,1]$$

$$\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{S}$$

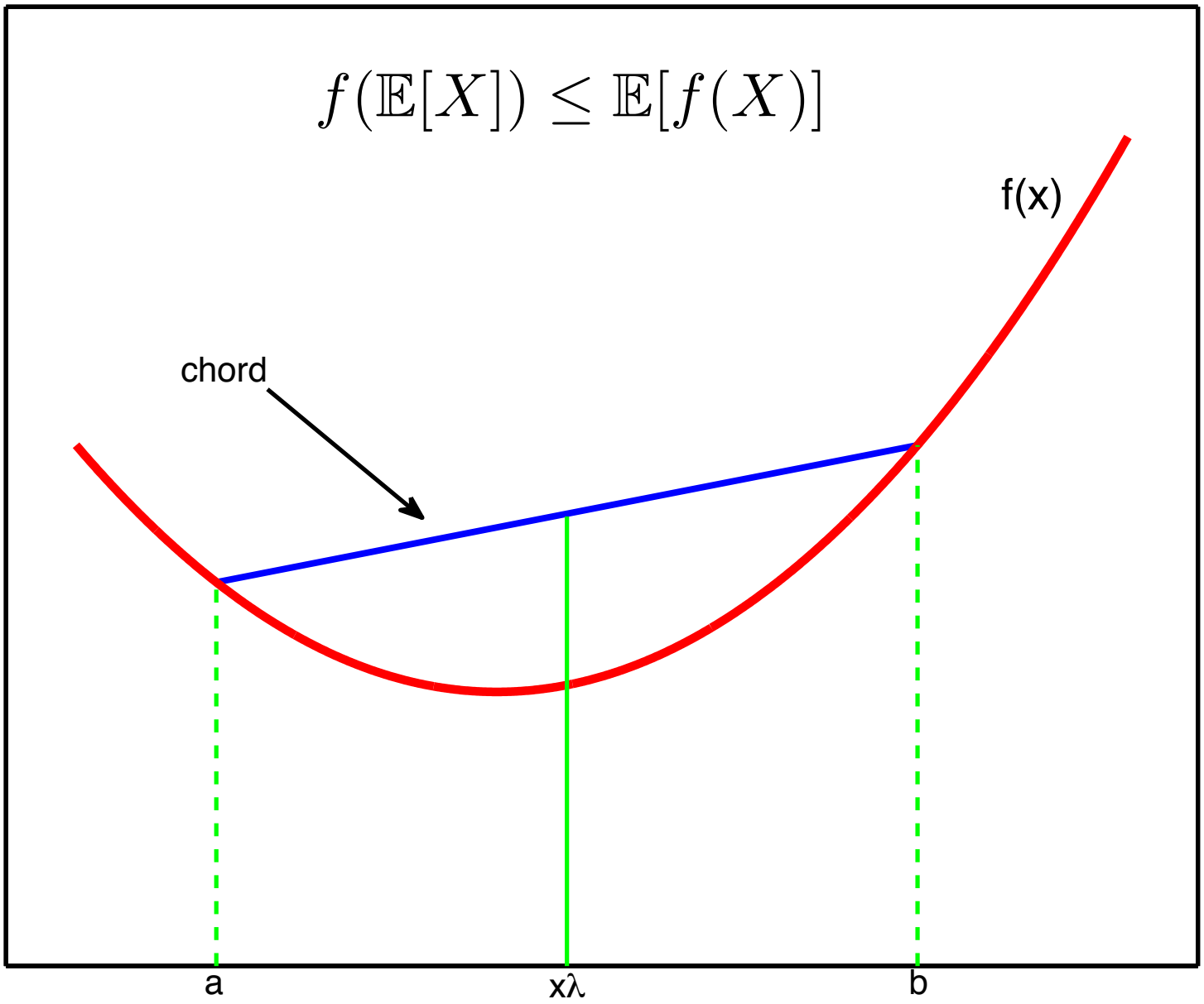






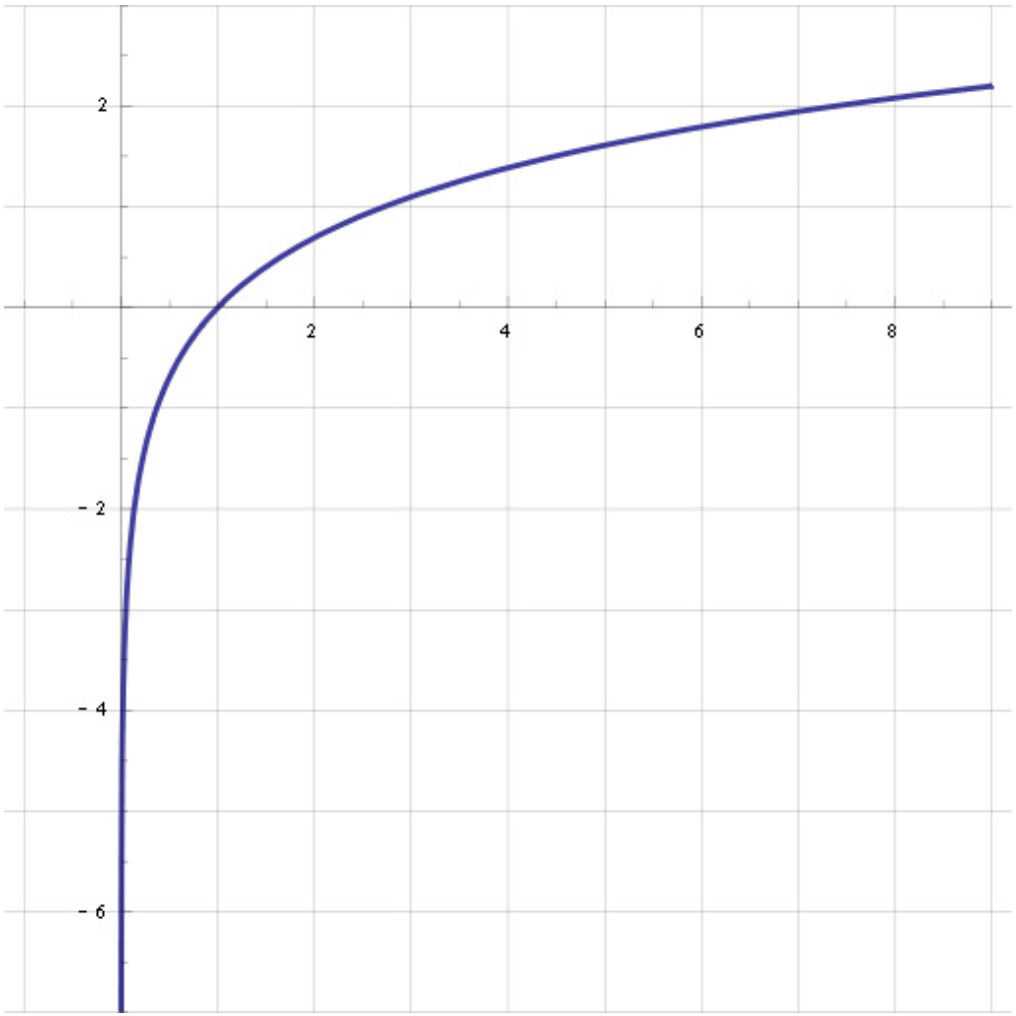
$$f(\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}') \leq \lambda f(\boldsymbol{\theta}) + (1-\lambda) f(\boldsymbol{\theta}')$$

Convexity & Jensen's Inequality



Concavity & Jensen's Inequality

$$\ln(\mathbb{E}[X]) \geq \mathbb{E}[\ln(X)]$$



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

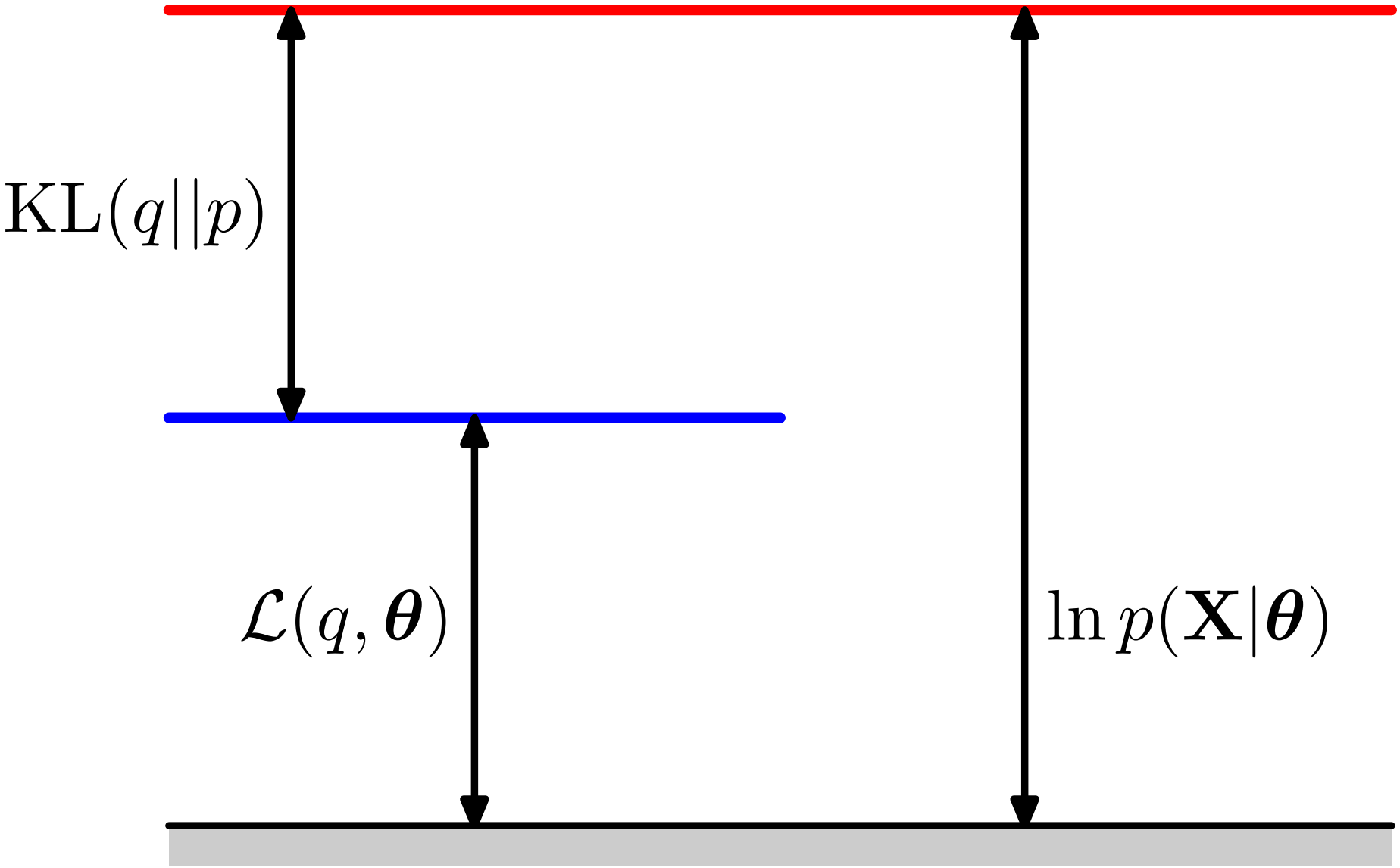
$$q^{(t)} = \arg\max_q \mathcal{L}(q, \theta^{(t-1)})$$

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

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

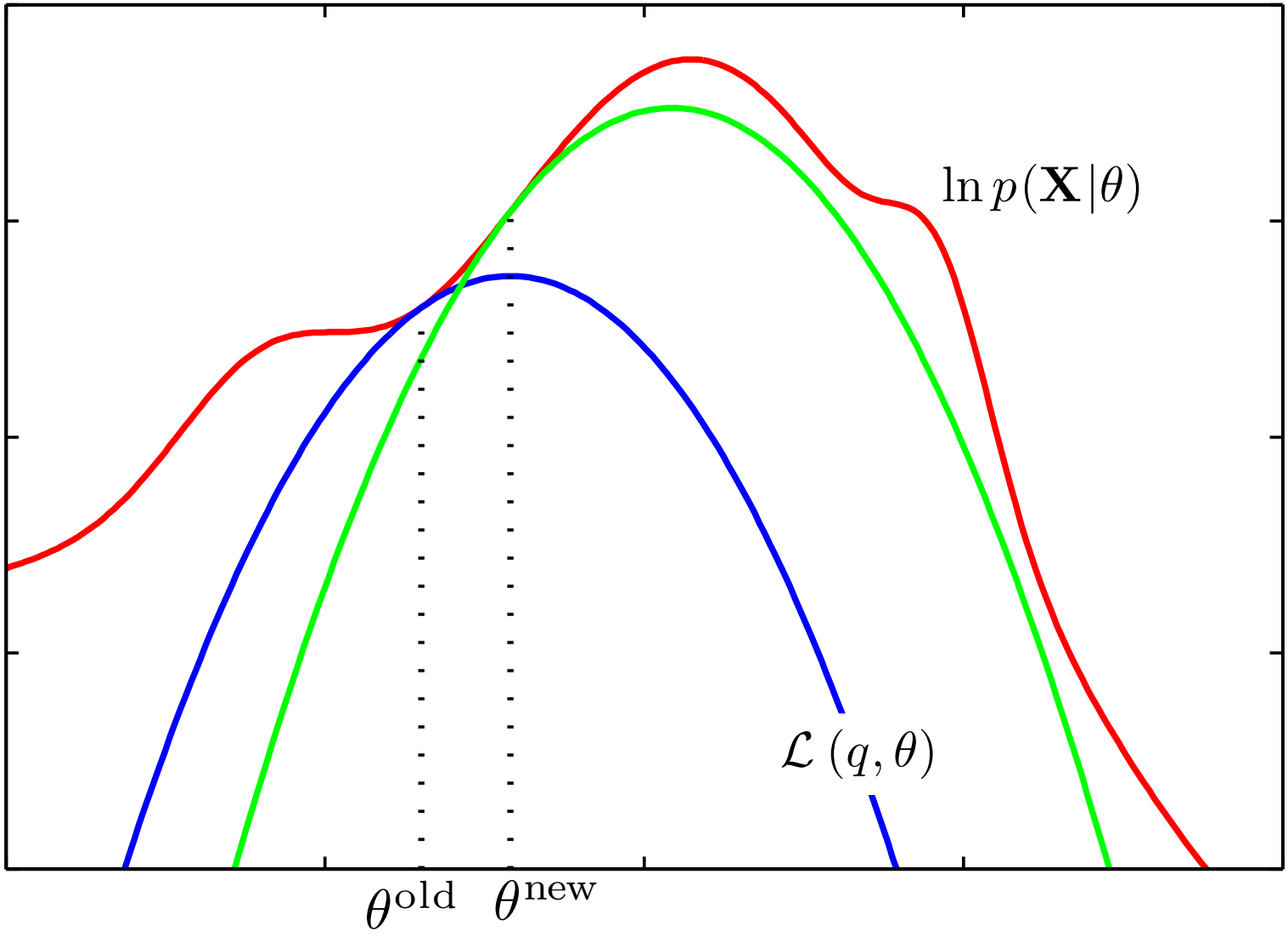
- **Iteration:** Alternate E-step & M-step until convergence

Lower Bounds on Marginal Likelihood



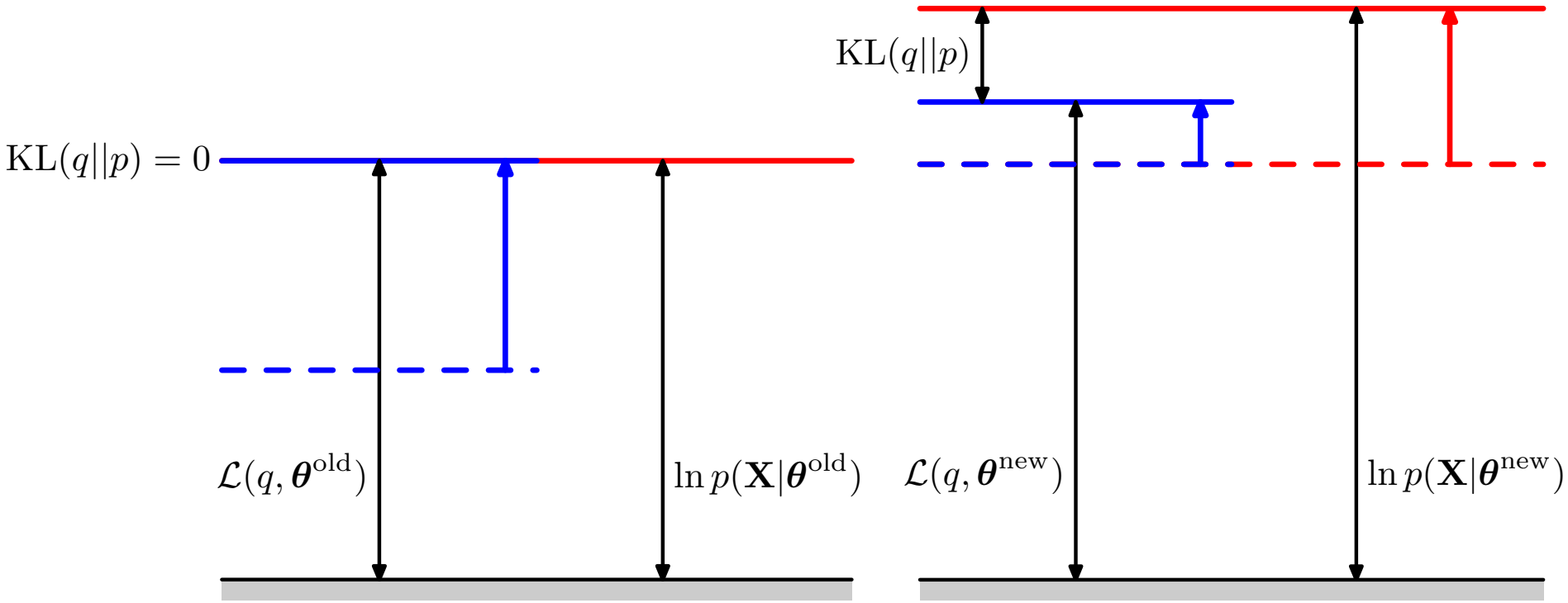
C. Bishop, Pattern Recognition & Machine Learning

EM: 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

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:

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

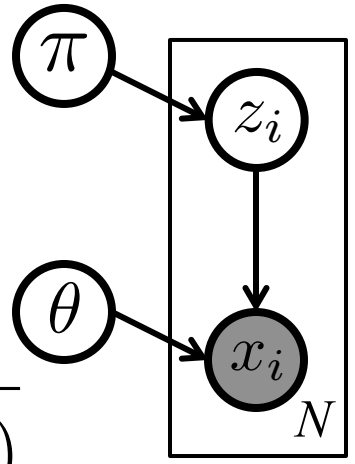
posterior distribution given current parameters

- Applying to probabilistic mixture models:

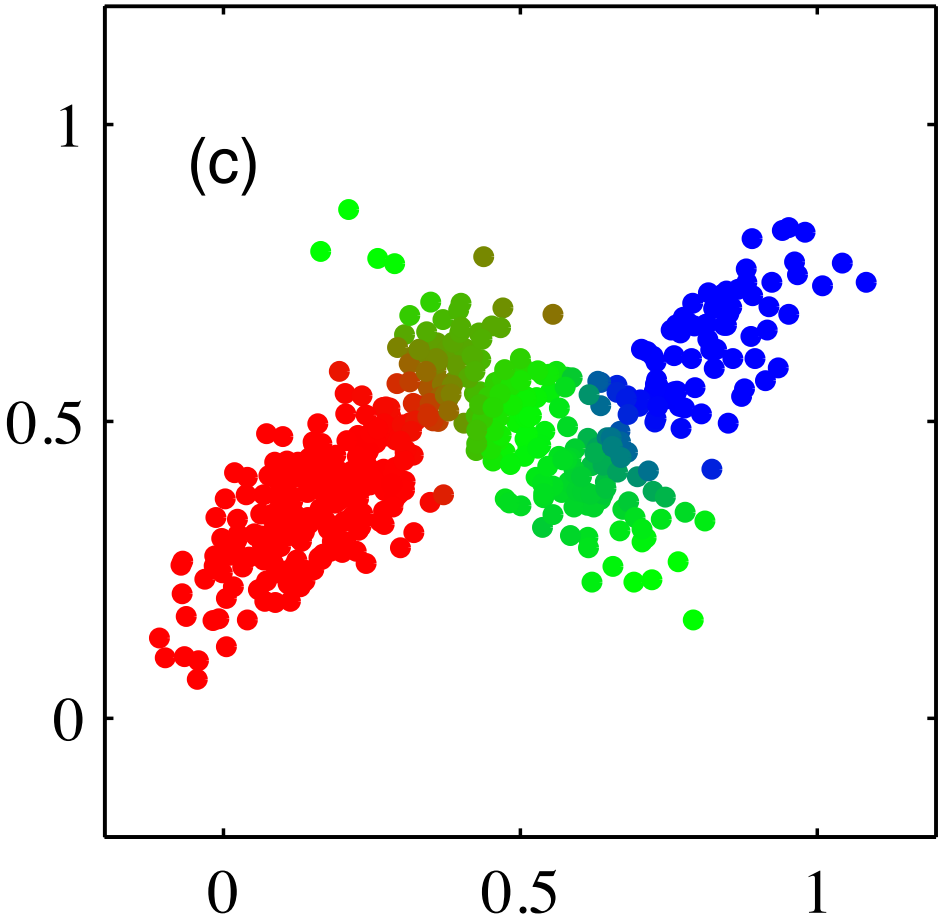
$$p(z_i \mid \pi) = \mathrm{Cat}(z_i \mid \pi)$$

$$p(x_i \mid z_i, \theta) = p(x_i \mid \theta_{z_i})$$

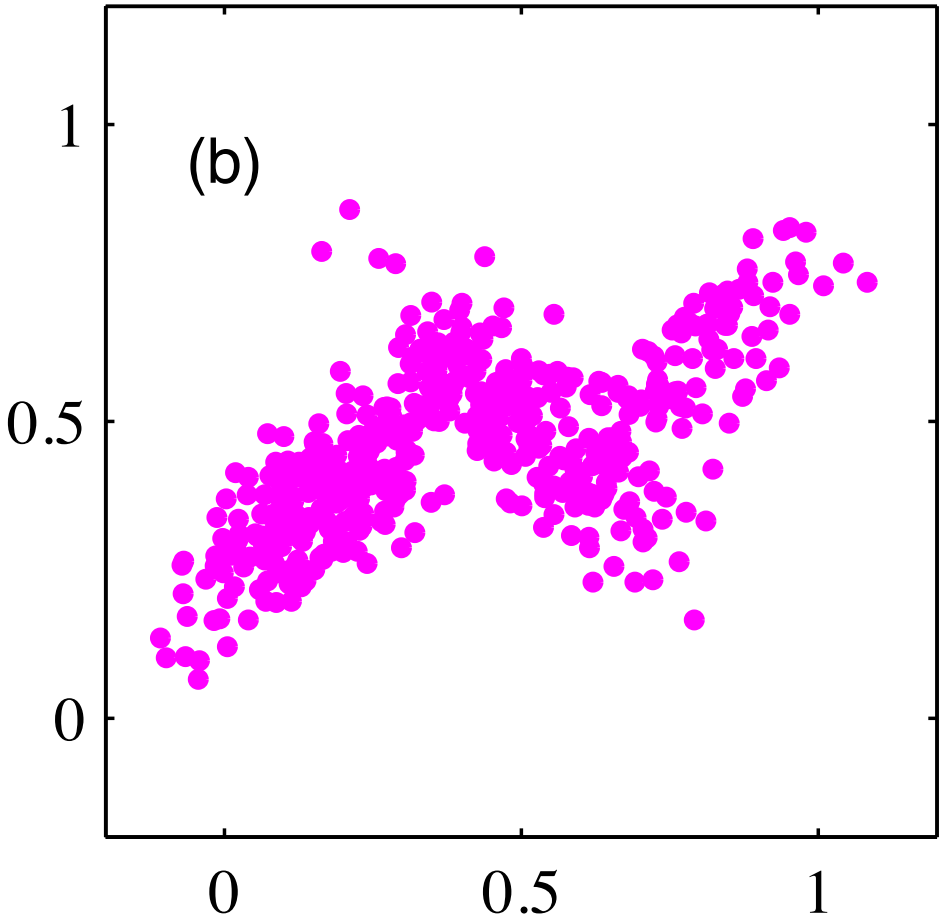
$$r_{ik} = p(z_i = k \mid x_i, \pi, \theta) = \frac{\pi_k p(x_i \mid \theta_k)}{\sum_{\ell=1}^{K} \pi_\ell p(x_i \mid \theta_\ell)}$$



E-Step for Gaussian Mixtures



Posterior Probabilities of Assignment to Each Cluster



Incomplete Data: Points to be Clustered

$$r_{ik} = p(z_i = k \mid x_i, \pi, \theta) = \frac{\pi_k p(x_i \mid \theta_k)}{\sum_{\ell=1}^{K} \pi_\ell p(x_i \mid \theta_\ell)}$$

Reminder: Exponential Families

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] = h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta})]$$

$$Z(\boldsymbol{\theta}) = \int_{\mathcal{X}^m} h(\mathbf{x}) \exp[\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x})] d\mathbf{x} \qquad A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$$

$\phi(x) \in \mathbb{R}^d \longrightarrow$ fixed vector of *sufficient statistics* (features), specifying the family of distributions

$\theta \in \Theta \longrightarrow$ unknown vector of *natural parameters*, determine particular distribution in this family

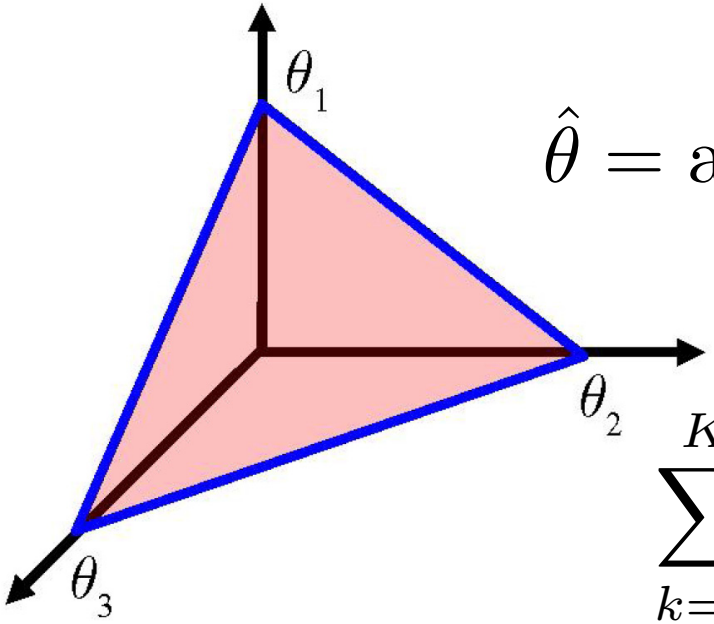
$Z(\theta) > 0 \longrightarrow$ normalization constant or *partition function*, ensuring this is a valid probability distribution

$h(x) > 0 \longrightarrow$ *reference measure* independent of parameters (for many models, we simply have $h(x) = 1$)

- Examples: Multinomial, Dirichlet, Gaussian, Poisson, …
- ML estimation via moment matching:

$$\mathbb{E}_\theta[\phi(x)] = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)$$

Reminder: Constrained Optimization



$$\hat{\theta} = \arg\max_{\theta} \sum_{k=1}^{K} a_k \log \theta_k \qquad a_k \geq 0$$

subject to

$$\sum_{k=1}^{K} \theta_k = 1 \qquad \theta_k \geq 0$$

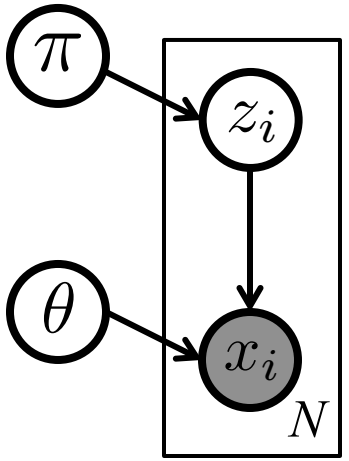
- Solution: $\hat{\theta}_k = \dfrac{a_k}{a_0} \qquad a_0 = \sum_{k=1}^{K} a_k$
- Proof for K=2: Change of variables to unconstrained problem
- Proof for general K: Lagrange multipliers (see textbook)

EM: Maximization Step

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

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

- Unlike E-step, no simplified general solution
- Applying to mixtures of *exponential families*:



$$p(z_i \mid \pi) = \mathrm{Cat}(z_i \mid \pi)$$

$$p(x_i \mid z_i, \theta) = \exp(\theta_{z_i}^T \phi(x_i) - A(\theta_{z_i}))$$

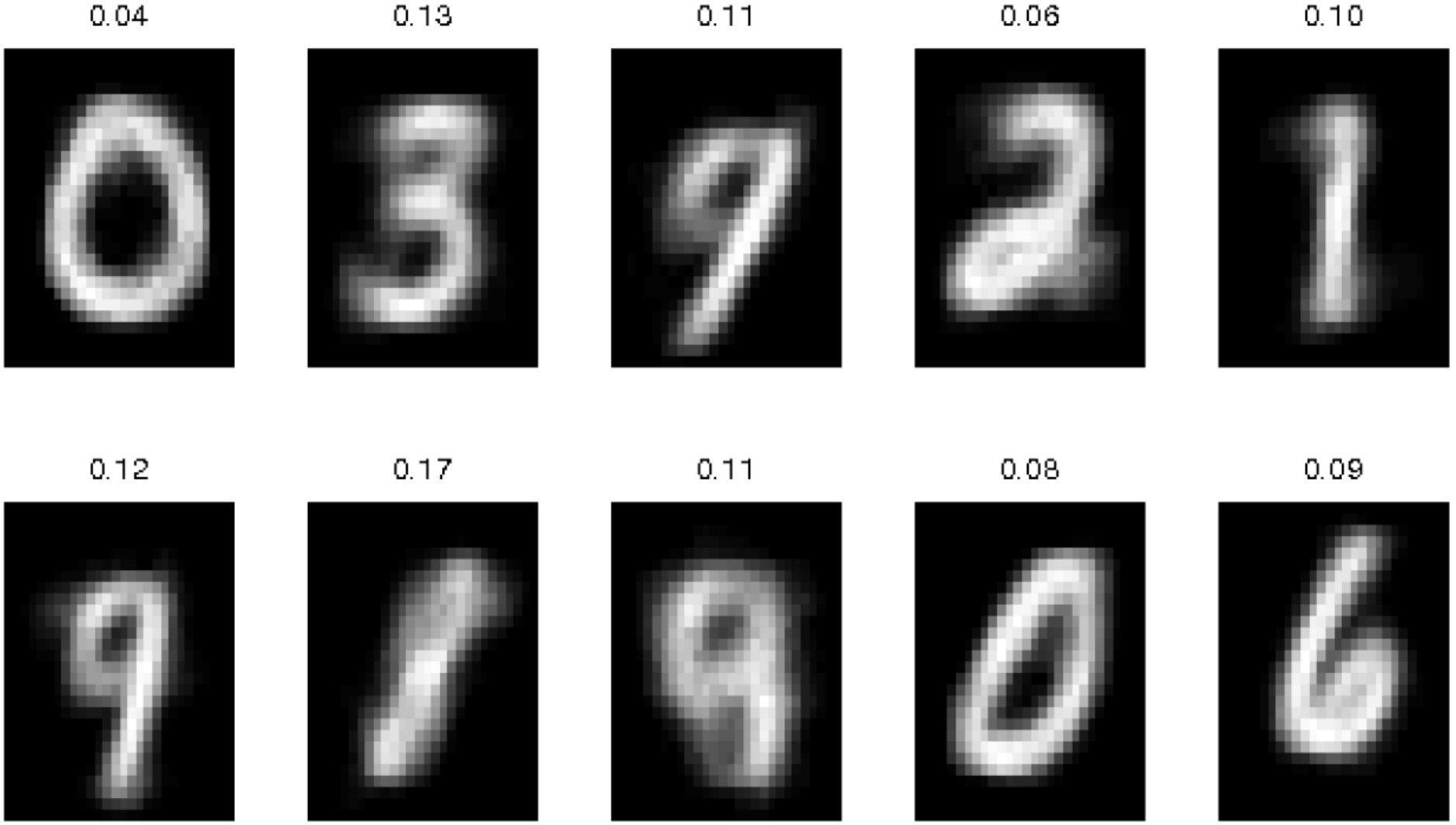
$$\hat{\pi}_k = \frac{N_k}{N}$$

$$\mathbb{E}_{\hat{\theta}_k}[\phi(x)] = \frac{1}{N_k} \sum_{i=1}^N r_{ik} \phi(x_i)$$

weighted moment matching

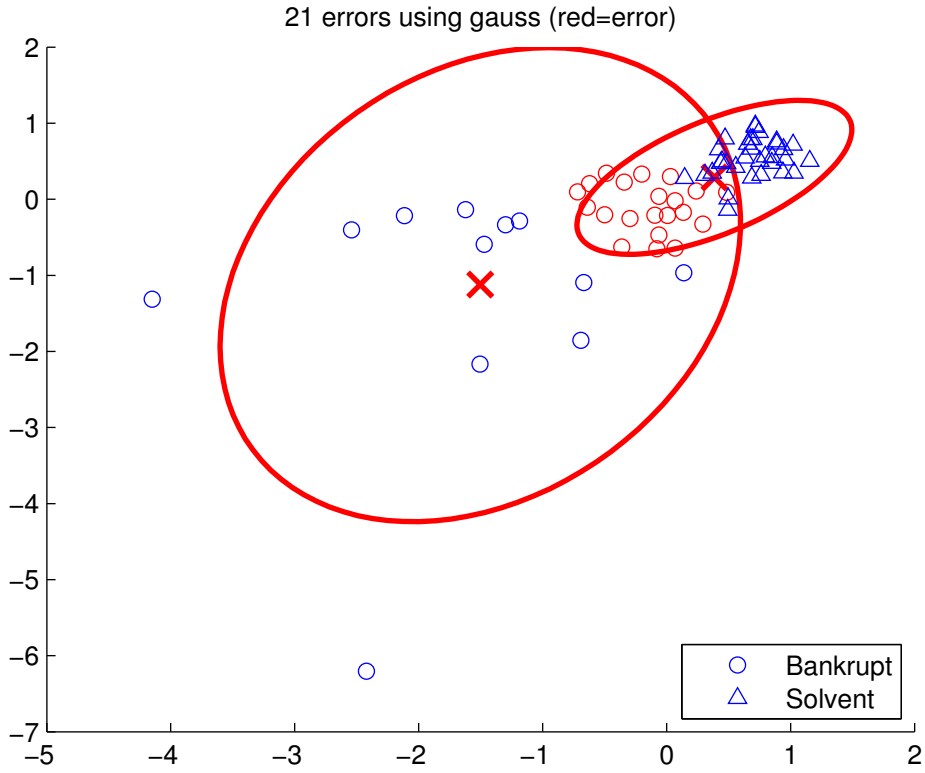
$$N_k = \sum_{i=1}^N r_{ik}$$

Binary Features: Mixtures of Bernoullis

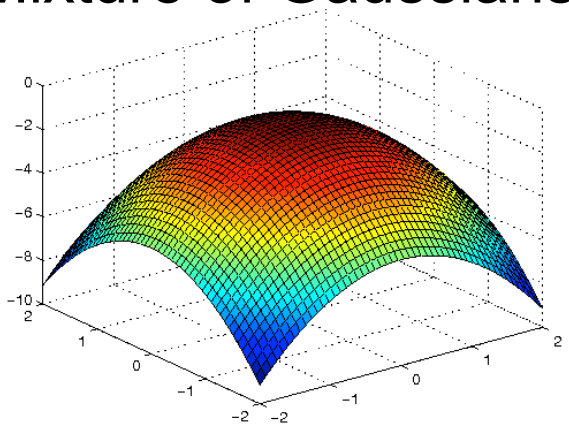


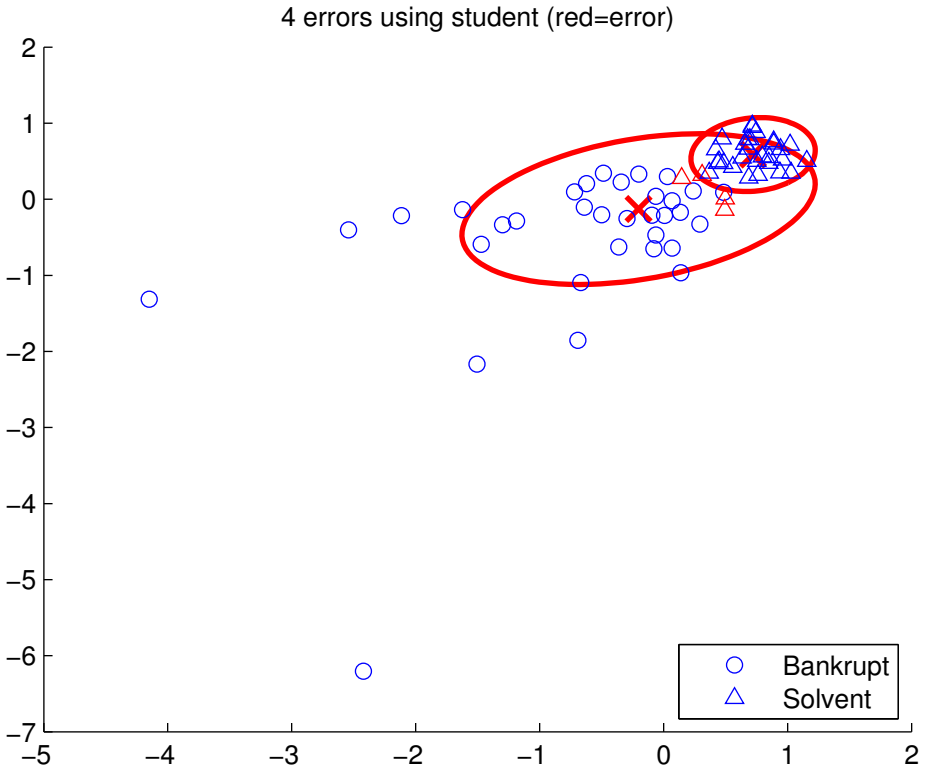
10 Clusters Identified via EM Algorithm from Binarized MNIST Digits

Density Shape Matters

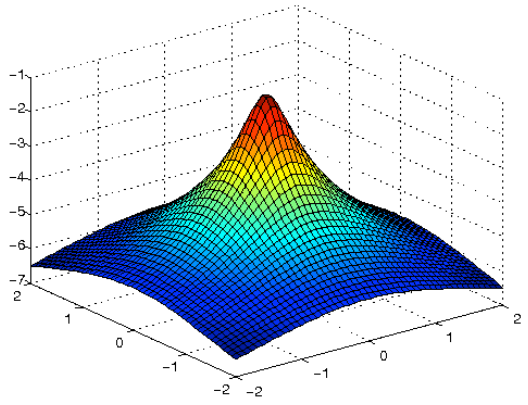


Mixture of Gaussians

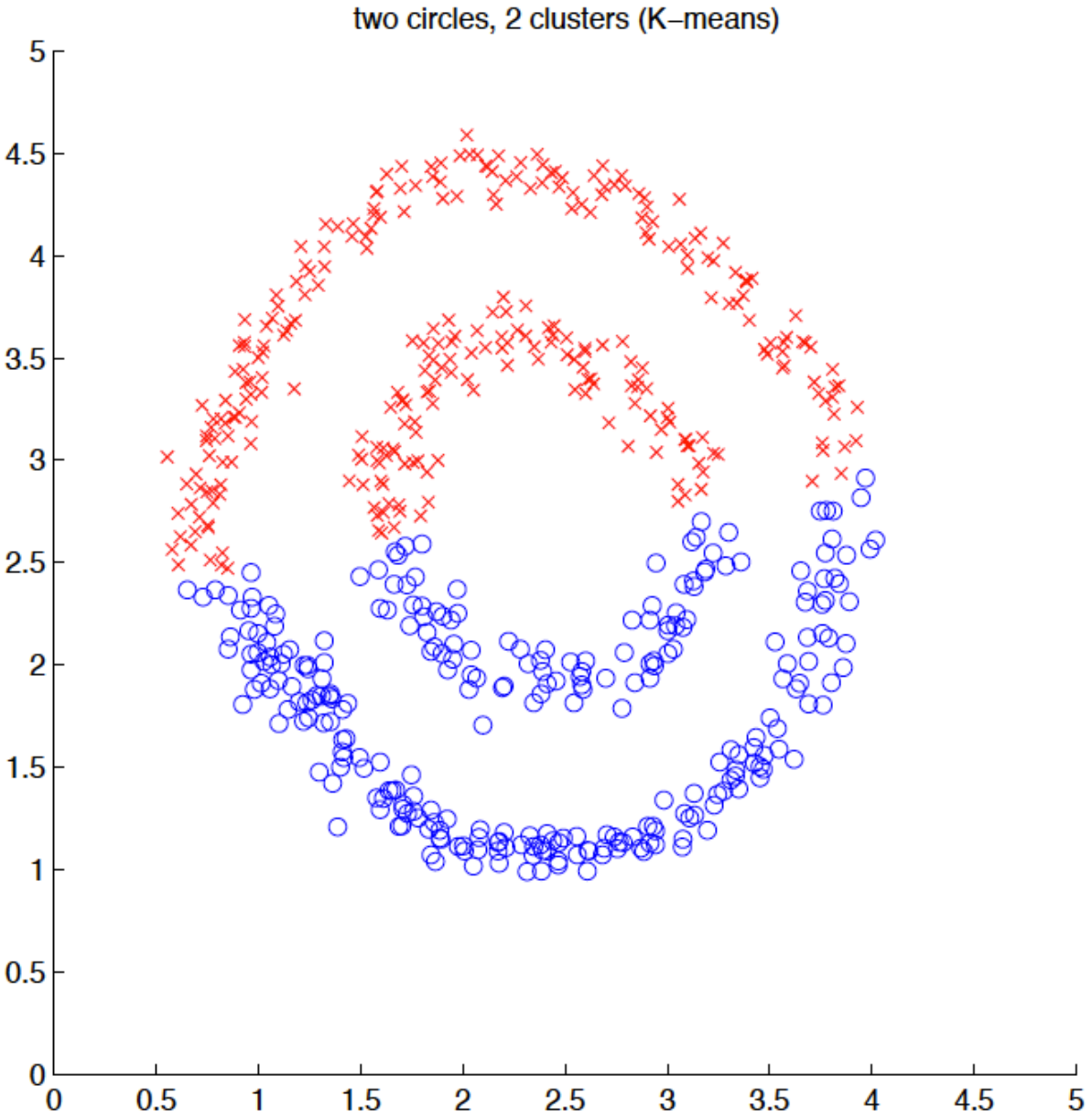




Mixture of Student t Densities



Real Clusters may not be Round



Ng, Jordan, & Weiss, NIPS02

Dimensionality Reduction

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

- **Goal:** Infer label/response y given only features x
- **Classical:** Find latent variables y good for *compression* of x
- **Probabilistic learning:** Estimate parameters of joint distribution p(x,y) which *maximize marginal probability* p(x)