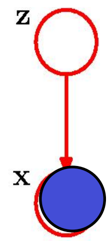
# Latent Variable View of EM

Sargur Srihari
srihari@cedar.buffalo.edu

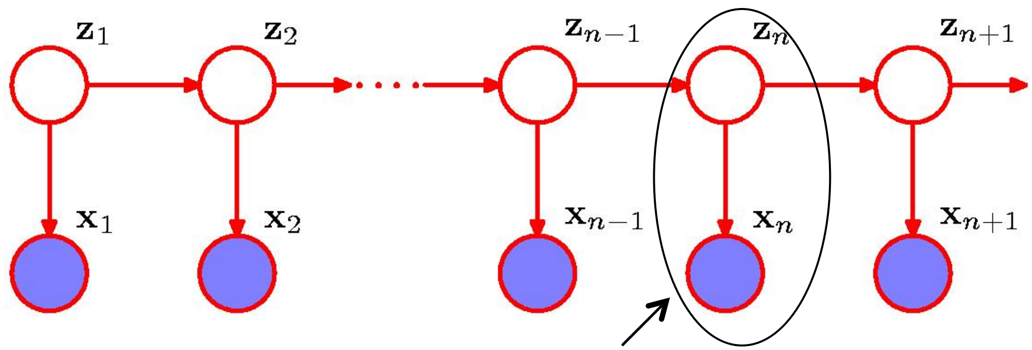
# Examples of latent variables

## 1. Mixture Model




- Joint distribution is $p(\mathrm{x},\mathrm{z})$
  - We don't have values for z

## 2. Hidden Markov Model

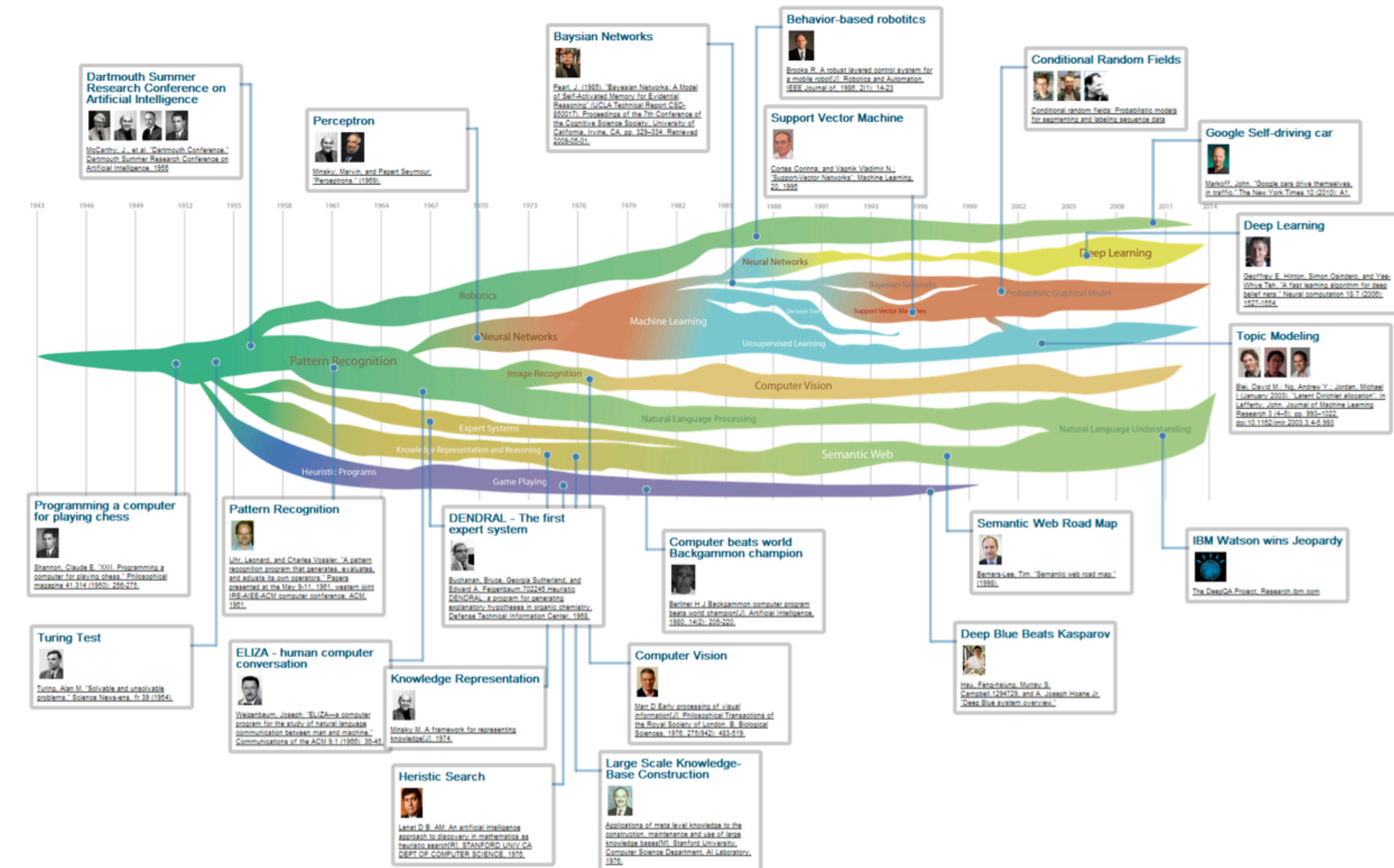



- A single time slice is a mixture with components $p(\mathrm{x}|\mathrm{z})$
- An extension of mixture model
  - Choice of mixture component depends on choice of mixture component for previous distribution
- Latent variables are multinomial variables $\mathrm{z}_n$
  - That describe component responsible for generating $\mathrm{x}_n$

# Another example of latent variables

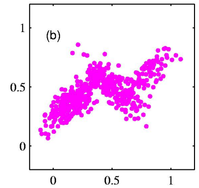
3. Topic Models (Latent Dirichlet Allocation)
   - In NLP unobserved groups explain why some observed data are similar
   - Each document is a mixture of various topics (latent variables)
   - Topics generate words
     - CAT-related: milk, meow, kitten
     - DOG-related: puppy, bark, bone
   - Multinomial distributions over words with Dirichlet priors

# ML as a subfield of AI



# Main Idea of EM

- Goal of EM is:
  - find maximum likelihood models for distributions $p(\mathrm{x})$ that have latent (or missing) data
    - E.g., GMMs, HMMs
- In case of Gaussian mixture models
  - We have a complex distribution of observed variables $\mathrm{x}$
  - We wish to estimate its parameters




- Introduce latent variables $\mathrm{z}$ so that
  - the joint distribution $p(\mathrm{x,z})$ of observed and latent variables is more tractable (since we know forms of components)
  - Complicated distribution is formed from simpler components
- The original distribution is obtained by marginalizing the joint distribution

# Alternative View of EM

- This view recognizes key role of latent variables
- Observed data $X = \begin{bmatrix} x_1 \\ x_2 \\ x_n \\ \end{bmatrix}$ Latent Variables $Z = \begin{bmatrix} z_1 \\ z_2 \\ z_n \\ \end{bmatrix}$
  - where $n^{th}$ row represents $x_n^T=[x_{n1}\ x_{n2}\ \ x_{nD}]$
  - with corresponding row $z_n^T=[z_{n1}\ z_{n2}\ \ z_{nK}]$
- Goal of EM algorithm is to find maximum likelihood solution for $p(X)$ given some $X$
- When we do not have $Z$

# Likelihood Function involving Latent Variables

- Joint likelihood function is $p(\mathrm{X},\mathrm{Z}\,|\,\theta)$ where $\theta$ is the set of all model parameters
  - E.g., means, covariances, responsibilities
- Marginal likelihood function of observed data
  - From sum rule

$$p(X\,|\,\theta)=\sum_{Z} p(X,Z\,|\,\theta)$$

- Log likelihood function is

$$\ln p(X\,|\,\theta)=\ln\left\{\sum_{Z} p(\mathrm{X},\mathrm{Z}\,|\,\theta)\right\}$$

# Latent Variables in EM

- Log likelihood function is

$$\ln p(X \mid \theta) = \ln\left\{\sum_{Z} p(X, Z \mid \theta)\right\}$$

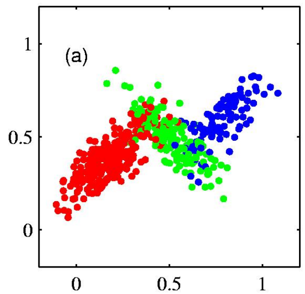
Summation inside brackets due to marginalization
Not due to log-likelihood

- Key Observation:
  - Summation over latent variables appears inside logarithm
    - Even if joint distribution $p(X, Z \mid \theta)$ belongs to exponential family the marginal distribution $p(X \mid \theta)$ does not
      - Taking log of Sum of Gaussians does not give simple quadratic
    - Results in complicated expressions for maximum likelihood solution, i.e., what value of $q$ maximizes the likelihood

# Complete and Incomplete Data Sets

## Complete Data $\{X,Z\}$

- For each observation in $X$ we know corresponding value of latent variable $Z$
- Log-likelihood has the form $p(X,Z \mid \theta)$
  - maximization is straightforward

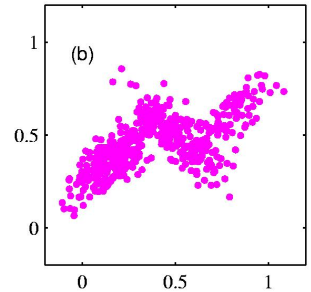



## Incomplete Data $\{X\}$

- Actual data set
- Log likelihood function is

$$\ln p(X \mid \theta) = \ln \left\{ \sum_{Z} p(X,Z \mid \theta) \right\}$$

- Maximization is difficult
  - summations inside logarithm



# Expectation of log-likelihood

- Since we don't have the complete data set $\{X,Z\}$ we evaluate the expected log-likelihood, i.e.,

$$E[\ln p(\mathrm{X},\mathrm{Z}|\theta)]$$

- Since we are given $X$, our state of knowledge of $Z$ is given only by the posterior distribution of the latent variables $p(\mathrm{Z} \mid \mathrm{X}, \theta)$
- Thus expected log-likelihood of complete data is

$$E\,[\,\ln p(X,Z \mid \theta)] = \sum_{Z} p(Z \mid X,\theta)\underbrace{\ln p(X,Z \mid \theta)}$$

Summation is due to expectation not sum rule!

We maximize this.
Note that the logarithm acts on the joint-- which is tractable

# E and M Steps

- *E Step*: Estimate the missing Values
  - Use current parameter value $\theta^{old}$ to find the posterior distribution of the latent variables given by

$$p(\mathrm{Z} \mid \mathrm{X}, \theta^{old})$$

- *M Step*: Determine revised parameter estimate $\theta^{new}$ by *maximizing* $\theta^{\mathrm{new}} = \arg\max_{\theta} Q(\theta, \theta^{\mathrm{old}})$
  - *where*

$$Q\left(\theta,\theta^{old}\right) = \sum_{Z} p\left(Z \mid X,\theta^{old}\right) p\left(X,Z \mid \theta\right)$$

Summation due to expectation

  - is the *expectation* of $p(X, Z \mid \theta)$ for some general parameter value $\theta$
  - Evaluate the log-likelihood $\sum_{i=1}^{N} \ln p(Xi, Z \mid \theta)$

# General EM Algorithm

- Given joint distribution $p(\mathrm{X},\mathrm{Z}\mid\theta)$ over observed variables $\mathrm{X}$ and latent variables $\mathrm{Z}$ governed by parameters $\theta$
  - goal is to maximize likelihood function $p(\mathrm{X}\mid\theta)$

- Step 1: Choose an initial setting for the parameters $\theta^{old}$
- Step 2: E Step: Evaluate $p(\mathrm{Z}\mid\mathrm{X},\theta^{old})$
- Step 3: M Step: Evaluate $\theta^{new}$ given by

$$\theta^{\text{new}} = \arg\max_{\theta} Q(\theta,\theta^{\text{old}})$$

where

$$Q(\theta,\theta^{old}) = \sum_{Z} p(Z\mid X,\theta^{old}) \ln p(X,Z\mid\theta)$$

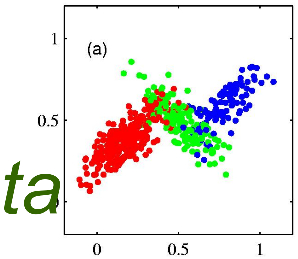
- Check for convergence
  - of either log-likelihood or parameter values
- If not satisfied then let $\theta^{old} \leftarrow \theta^{new}$
- Return to Step 2

# Missing Variables

- EM has been described for maximum likelihood function when there are discrete latent variables
- It can also be applied when there are unobserved variables corresponding to missing values in data set
  - Take the joint distribution of all variables and then marginalize over missing ones
  - EM is then used to maximize corresponding likelihood function
- Method is valid when data is missing at random
  - Not if missing value depends on unobserved values
  - E.g., if quantity exceeds some threshold

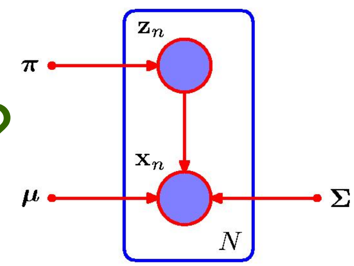
# Gaussian Mixtures Revisited

- Apply EM (latent variable view) to GMM
- In the E-step we compute
  - Expectation of *log-likelihood of complete data* $\{X,Z\}$ wrt posterior of latent Variables $Z$




$$\boxed{Q(\theta,\theta^{old}) = \sum_{Z} p(Z \mid X,\, \theta^{old}) \ln p(X,Z \mid \theta)}$$

  - What is the form of the two product terms?




- In the M-step we maximize $Q(\theta,\theta^{old})$ wrt $\theta$
  - Will show that this leads to the same m.l estimates for GMM parameters $\pi, \mu, \Sigma$ as before

# Likelihood for Complete Data

- Likelihood function for the complete data set is

$$p(X,Z|\pi,\mu,\Sigma)=\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}N\left(x_n|\mu_k,\Sigma_k\right)^{z_{nk}}$$

- Log-likelihood is

$$\ln p(X,Z|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\left\{\ln\pi_k+\ln N\left(x_n|\mu_k,\Sigma_k\right)\right\}$$

- Much simpler than log-likelihood for incomplete data:

$$\ln p(X|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k N(\mathrm{x}_n|\mu_k,\Sigma_k)\right\}$$

- Maximum likelihood solution for complete data can be obtained in closed form
- Since we don't have values for latent variables, we obtain its expectation wrt the posterior distribution of latent variables

# Posterior Distribution of Latent Variables

- From $p(\mathrm{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$ and $p(\mathrm{x} \mid \mathrm{z}) = \prod_{k=1}^{K} N(x \mid \mu_k, \Sigma_k)^{z_k}$ we have

$$\boxed{p(Z \mid X, \mu, \Sigma) \; \alpha \; \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k N\left(x_n \mid \mu_k, \Sigma_k\right)\right)^{z_{nk}}}$$

- From which we can get the expected value for the indicator variable as

$$\boxed{E[z_{nk}] = \frac{\pi_k N\left(x_n \mid \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \pi_j N\left(x_n \mid \mu_j, \Sigma_j\right)} = \gamma\left(z_{nk}\right)}$$

- Substituting into complete log-likelihood:

$$\boxed{E_Z\left[\ln p(X, Z \mid \pi, \mu, \Sigma)\right] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma\left(z_{nk}\right)\left\{\ln \pi_k + \ln N\left(x_n \mid \mu_k, \Sigma_k\right)\right\}}$$

- Final procedure: choose initial values for $\pi^{old}, \mu^{old}, \Sigma^{old}$
  - Evaluate the responsibilities (E-step)
  - Keep responsibilities fixed and use closed-form solutions for $\pi^{new}, \mu^{new}, \Sigma^{new}$

$$\boxed{\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathrm{x}_n} \qquad \boxed{\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathrm{x}_n - \mu_k)(\mathrm{x}_n - \mu_k)^T} \qquad \boxed{\pi_k = \frac{N_k}{N}}$$

# Relation to K-means

- EM for Gaussian mixtures has close similarity to K-means
- K-means performs a hard assignment of data points to clusters
  - Each data point is associated uniquely with one cluster
- EM makes a soft assignment based on posterior probabilities
- K-means does not estimate the covariances of the clusters but only the cluster means

# Mixtures of Bernoulli Distributions

- Previously considered distributions over continuous variables
- Now consider mixtures of discrete binary variables described by Bernoulli distributions
- Sets a foundation for HMMs over discrete variables

# Multivariate Bernoulli

- Set of $D$ independent binary variables $x_i$, $i=1,..,D$
  - E.g., a set of $D$ coins with heads and tails
- Each governed by parameter $m_i$
- Multivariate distribution

$$p(\mathrm{x}|\mu) = \prod_{i=1}^{D} \mu_i^{x_i} (1-\mu_i)^{(1-x_i)}$$

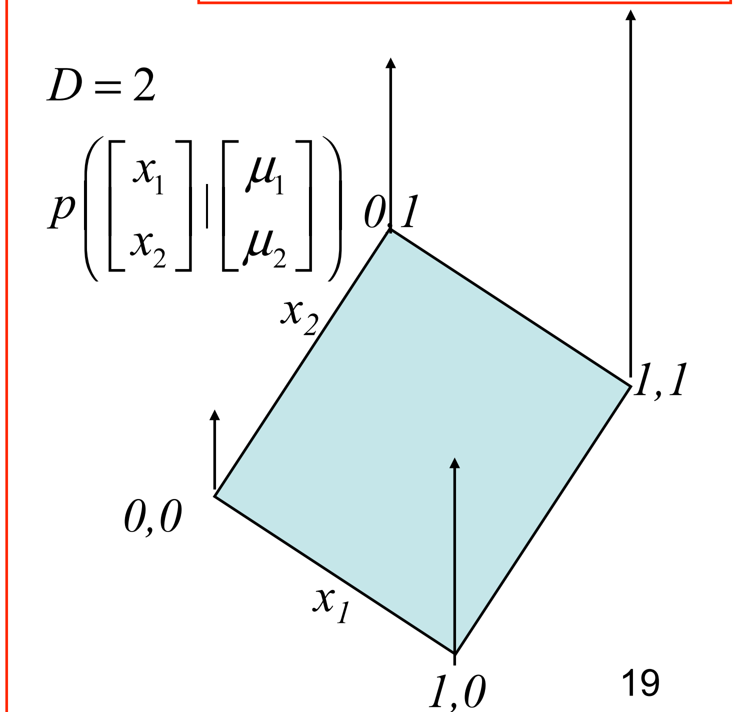
where $\mathrm{x}=(x_1,..,x_D)^T$ and

$\mathrm{m} =(m_1,..,m_D)^T$

- Mean and covariance are

$\mathrm{E[x]=m},\ \mathrm{cov[x]=diag}\{m_i(1-m_i)\}$

A Bernoulli distribution with two variables



# Mixture of multivariate Bernoulli

- Finite mixture of $K$ Bernoulli distributions
  - E.g., $K$ bags of $D$ coins each where bag $k$ is chosen with probability $\mathbf{p}_k$

$$p(\mathrm{x} \mid \mu, \pi) = \sum_{k=1}^{K} \pi_k \, p(\mathrm{x} | \mu_k)$$

  - Where $\mathbf{m}=\{\mathbf{m}_1,..,\mathbf{m}_K\}$, $\mathbf{p}=\{\mathbf{p}_1,..,\mathbf{p}_K\}$, and

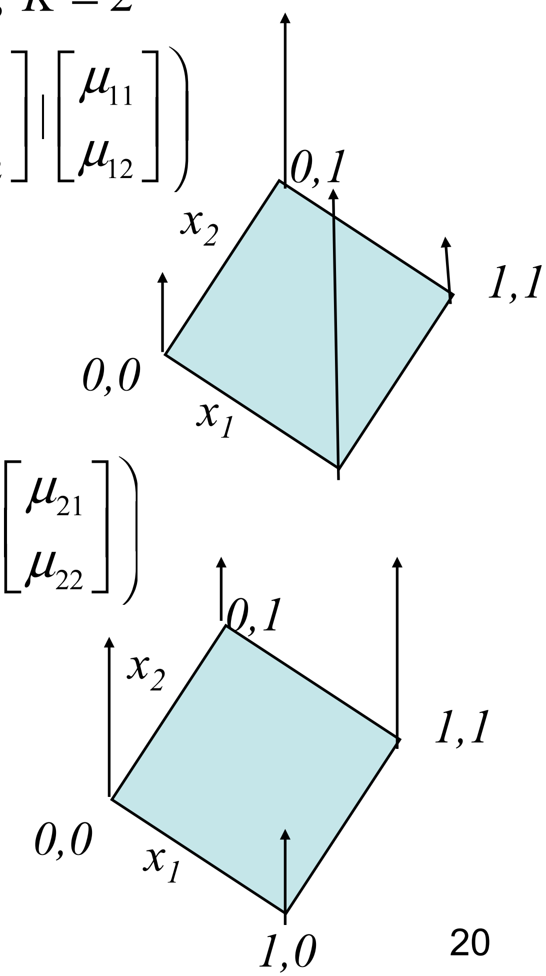
$$p(\mathrm{x}|\mu_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1-\mu_{ki})^{(1-x_i)}$$

Two Bernoulli distributions with two variables

$D = 2,\ K = 2$

$$p\left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} \mu_{11} \\ \mu_{12} \end{bmatrix} \right)$$

$$p\left( \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} \mu_{21} \\ \mu_{22} \end{bmatrix} \right)$$



# Log likelihood of Bernoulli mixture

- Given data set $X=\{x_1,..,x_N\}$ log likelihood of model is

$$\ln p(X \mid \mu,\pi) = \underbrace{\sum_{n=1}^{N}}_{\text{Summation due to logarithm}} \ln \left\{ \underbrace{\sum_{k=1}^{K} \pi_k p(x_n \mid \mu_k)}_{\text{Summation due to mixture}} \right\}$$

- Due to summation inside logarithm there is no closed form m.l.e. solution

# Introduce latent variables

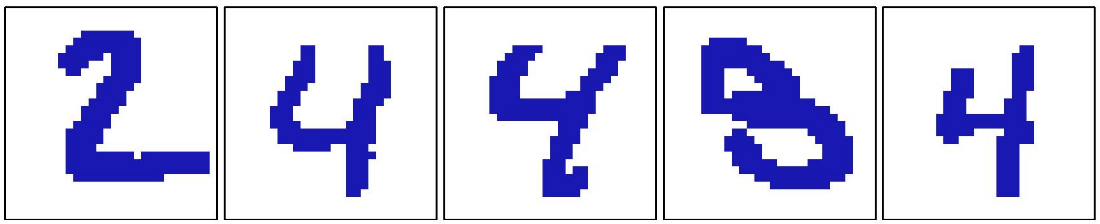
- One of $K$ representation $\mathrm{z} = (\mathrm{z}_1,..\mathrm{z}_\mathrm{K})^\mathrm{T}$
- Conditional distribution of $\mathrm{x}$ given the latent variable is

$$p(\mathrm{x}|\mathrm{z},\mu) = \prod_{k=1}^{K} p(\mathrm{x}|\mu_k)^{z_k}$$

- The EM algorithm is derived by writing the complete data log-likelihood function
- Taking its expectation w.r.t. posterior distribution of the latent variables
- In the E step these responsibilities are evaluated using Bayes theorem
- In the M step we maximize the expected complete data loglikelihood wrt parameters m$_k$ and p

# Illustration of Bernoulli Mixture

We are given a set of unlabeled digits 2,3 and 4
Goal is to use EM to cluster them with $K=3$

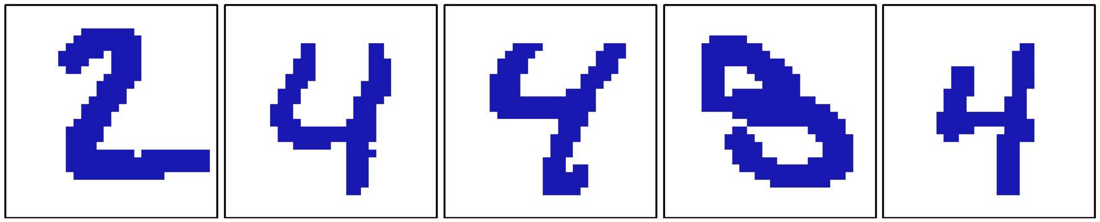


Binary images after grey-scale thresholding at 0.5
$N = 600$

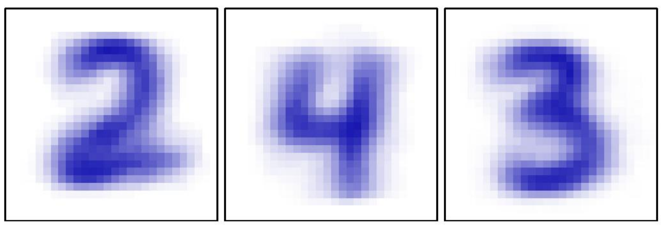
Mixing coefficients initialized with $p_k=1/K$

Parameters $m_{ki}$ were set to random values chosen uniformly in range $(0.25,0.75)$ and normalized so that $S_j m_{kj}=1$

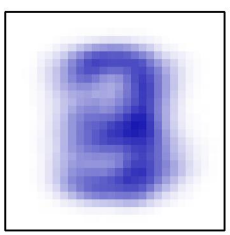
# Result of EM algorithm



EM finds the three clusters



Parameters $m_{ki}$ for each of three components of mixture model



Using single multivariate Bernoulli and maximum likelihood amounts to averaging counts in each pixel

# Bayesian EM for Discrete Case

- Conjugate prior of the parameters of Bernoulli is given by the beta distribution
- Beta prior is equivalent to introducing additional effective observations of $\mathbf{x}$
- Also introduce priors into the Bernoulli mixture model
- Use EM to maximize posterior probability of distribution
- Can be extended to multinomial discrete variables
  - Introduce Dirichlet priors over model parameters if desired

# EM Algorithm in General

- EM is a general technique for finding maximum likelihood solutions for probabilistic models with latent variables (Dempster 1977)
- EM defined heuristically can be proved to maximize the likelihood function
- Proof involves obtaining lower bound on log- likelihood function

# Proof for EM Algorithm

- Given observed variables X and hidden variables Z, goal is to maximize likelihood function

$$p(X \mid \theta) = \sum_Z p(X, Z \mid \theta)$$

- Direct optimization of $p(X|q)$ is difficult, whereas complete-data likelihood $p(X,Z|q)$ is easier (since component forms are tractable)
- For any choice of $q(Z)$ the following holds

$$\ln p(X \mid \theta) = L(q, \theta) + KL(q \,\|\, p)$$

where we define

$$\mathrm{L}(\mathrm{q}, \theta) = \sum_{\mathrm{z}} \mathrm{q}(\mathrm{Z}) \ln \left\{ \frac{\mathrm{p}(\mathrm{X}, \mathrm{Z} \mid \theta)}{\mathrm{q}(\mathrm{Z})} \right\}$$

$$\mathrm{KL}(\mathrm{q} \,\|\, \mathrm{p}) = -\sum_{\mathrm{z}} \mathrm{q}(\mathrm{Z}) \ln \left\{ \frac{\mathrm{p}(\mathrm{Z} \mid \mathrm{X}, \theta)}{\mathrm{q}(\mathrm{Z})} \right\}$$

This decomposition uses:

$$\ln p(X, Z \mid \theta) = \ln p(Z \mid X, \theta) + \ln p(X \mid \theta)$$

L is a *functional* that takes a function as input and produces a value as output, like entropy
It contains joint distribution of $X$ and $Z$

KL is the Kullback-Leibler Divergence
Contains conditional distribution of $Z$ given $X$

# Bounds on the log-likelihood

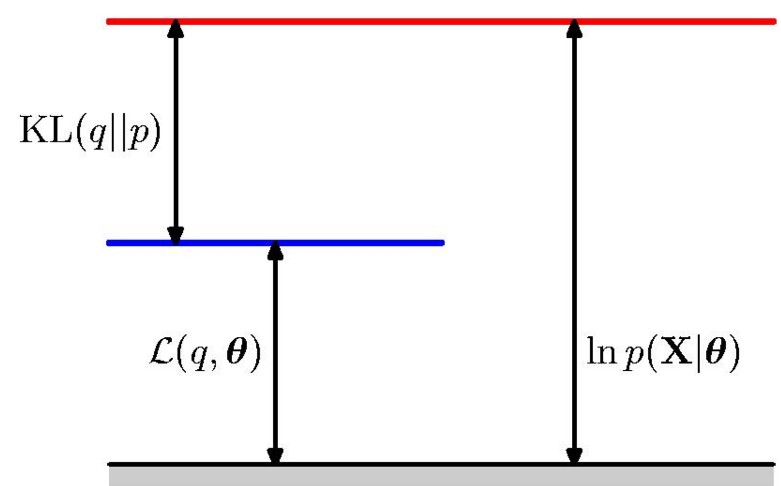
- Decomposition used to show that EM finds Maximum likelihood solution for $p(X/\mathrm{q})$

$$\ln p(X \mid \theta) = L(q,\theta) + KL(q \parallel p) \quad \text{where we define}$$

$$\mathrm{L}(\mathrm{q},\theta) = \sum_{\mathrm{z}} \mathrm{q}(\mathrm{Z}) \ln \left\{ \frac{\mathrm{p}(\mathrm{X},\mathrm{Z} \mid \theta)}{\mathrm{q}(\mathrm{Z})} \right\} \qquad \mathrm{KL}(\mathrm{q} \parallel \mathrm{p}) = -\sum_{\mathrm{z}} \mathrm{q}(\mathrm{Z}) \ln \left\{ \frac{\mathrm{p}(\mathrm{Z} \mid \mathrm{X},\theta)}{\mathrm{q}(\mathrm{Z})} \right\}$$

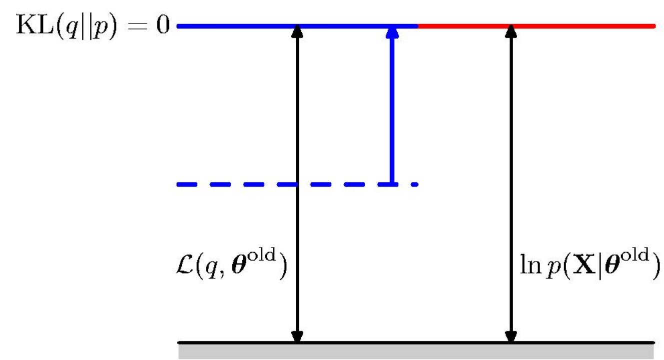
*L* is the lower bound

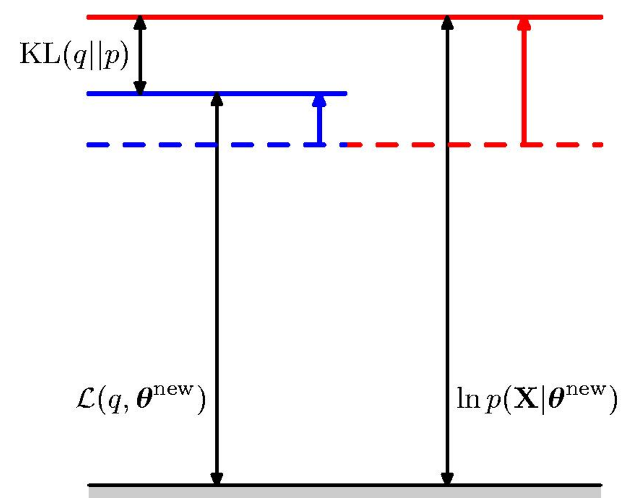
- Since $\mathrm{KL}(\mathrm{q}||\mathrm{p}) \geq 0$ It follows that $\mathrm{L}(\mathrm{q},\mathrm{q}) \leq \ln p(\mathrm{X}|\mathrm{q})$
  - Or $\mathrm{L}(\mathrm{q},\mathrm{q})$ is a lower bound on $\ln \mathrm{p}(\mathrm{X}|\mathrm{q})$



# Illustration of E and M steps

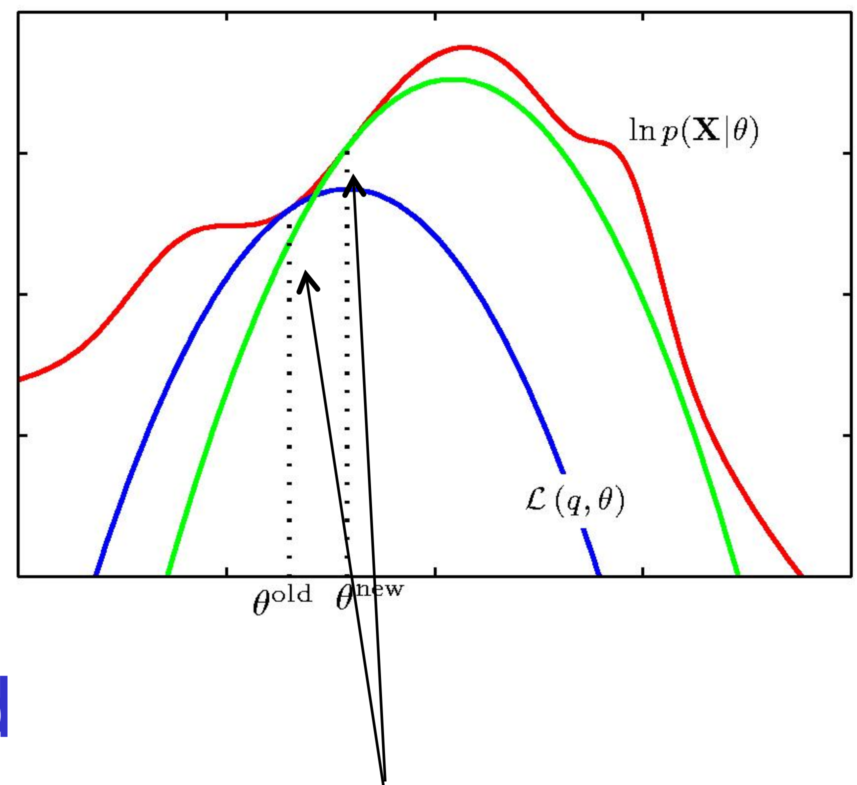
- E step (Lower bound maximized keeping $q^{old}$ fixed)
  - $q$ distribution is set to posterior distribution for current parameter values $q^{old}$
  - Causing lower bound to move to same value as log-likelihood with KL vanishing
- M step
  - Lower bound is maximized wrt $q$ to give revised value $q^{new}$
  - Causes log-likelihood to increase by as much as lower bound



# View of EM in parameter space

- EM involves alternately computing lower bound on log-likelihood for the current parameter values
- And maximizing this bound to obtain new parameter values
- Note that the lower bound is a convex function with a unique maximum




Lower bound on log-likelihood function tangential at $q^{old}$ and $q^{new}$

# Generalized EM (GEM)

- EM breaks own potentially difficult problem of maximizing the likelihood function into two stages, the E step and the M step
- One or both may remain intractable

- GEM addresses the problem of the intractable M step
  - Instead of maximizing $L(q,\mathbf{q})$ wrt $\mathbf{q}$ it changes parameters so as to increase its value
- Similar generalization of the E step can be made