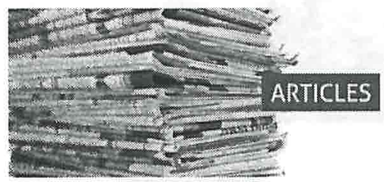
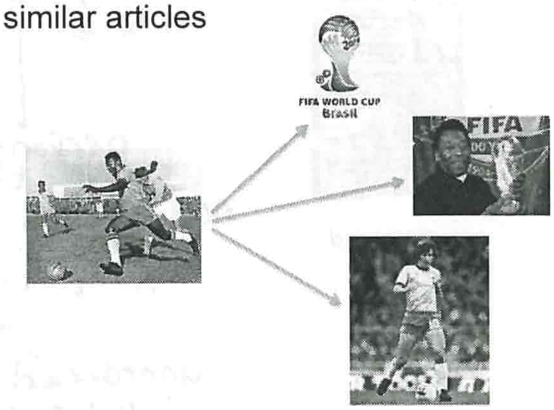
Document Retrieval

- **Goal:** Retrieve documents of interest
- **Challenges:**
  - ☐ Tons of articles out there
  - ☐ How should we measure similarity?



Task 1: Find Similar Documents

- **So far…**
  - ☐ **Input:** Query article x
  - ☐ **Output:** Set of k similar articles

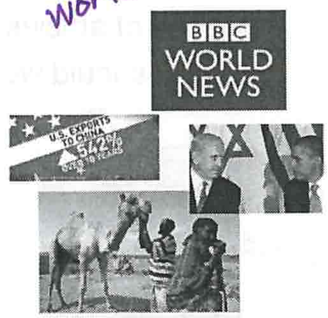


Task 2: Cluster Documents

- **Now:**
  - Cluster documents based on topic

sports

world news

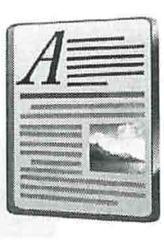


~~Many motivations including doc retrieval~~

More global description of corpus

Document Representation

- Bag of words model



document *d*

previously: $x = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$ ← vector fcn of word counts (e.g. tf-idf)

performed operations on this vector

now:

$x = \{w_1, \dots, w_N\}$

unordered set of N words $w_i \in V$ (vocab) in doc

A Generative Model

probabilistic model for simulating obs.

- Documents: $x^1, \ldots, x^D$ with $x^d = \{w_1^d, \ldots, w_{N_d}^d\}$
- Associated topics: $z^1, \ldots, z^D$ with $z^d \in \{1, \ldots, K\}$
- Parameters: $\theta = \{\pi, \beta\}$

total # of topics / clusters

$\pi = [\pi_1, \ldots, \pi_K]$ topic probabilities

with $Pr(z^d = k) = \pi_k$

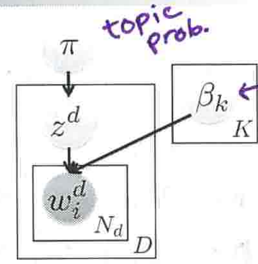
words

$$\beta = \text{topics} \begin{array}{c|ccc} & 1 & \cdots & V \\ \hline 1 & & \beta_1 & \\ \vdots & & \vdots & \\ K & & \beta_K & \end{array}$$

word probabilities for each topic

A Generative Model

- Documents: $x^1, \ldots, x^D$
- Associated topics: $z^1, \ldots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:



topic prob.

word prob. for topic k

"drawn from"

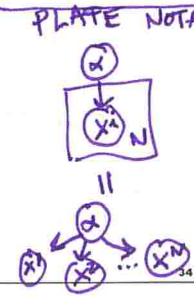
$$z^d \sim \pi$$

$$w_i^d \mid z^d \sim \beta_{z_d} \quad i = 1, \ldots, N_d$$

"given"

Given topic $z^d = k$ for doc d, draw each word ind. from β_k

PLATE NOTATION



Form of Likelihood

- Conditioned on topic...

$$p(x^d \mid z^d, \beta) = \prod_{i=1}^{N_d} p(w_i^d \mid z^d, \beta) = \prod_{i=1}^{N_d} \beta_{z^d w_i^d}$$

unobserved/latent

- Marginalizing latent topic assignment:

$$p(x^d \mid \beta, \pi) = \sum_{k=1}^{K} p(x^d, z^d = k \mid \beta, \pi)$$

$$= \sum_{k=1}^{K} p(x^d \mid z^d = k, \beta)\, p(z^d = k \mid \pi)$$

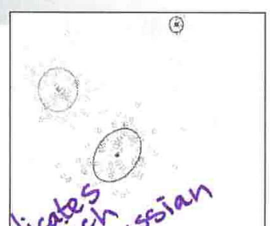
convex comb. of $p(x^d \mid z^d, \beta)$

$$= \sum_{k=1}^{K} \pi_k\, p(x^d \mid z^d = k, \beta)$$

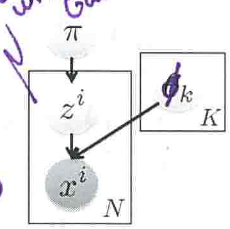
Gaussian Mixture Model

- Most commonly used mixture model
- Observations: $x^1, \ldots, x^N$
  with $x^i \in \mathbb{R}^d$
- Parameters:
  $\pi = [\pi_1, \ldots, \pi_K]$
  $\phi = \{\phi_k\} = \{\mu_k, \Sigma_k\}$
  params for cluster k
- Likelihood:

$$p(x^i \mid \theta) = \sum_{k=1}^{K} \pi_k\, N(x^i; \mu_k, \Sigma_k)$$



indicates which Gaussian

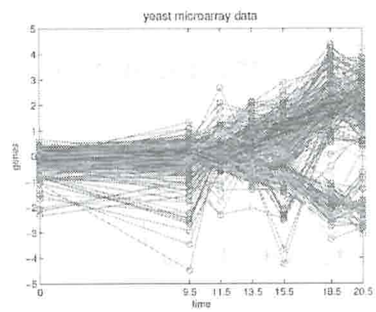


- Ex. z^i = country of origin, x^i = height of i[th] person
  - k^{th} mixture component = distribution of heights in country k

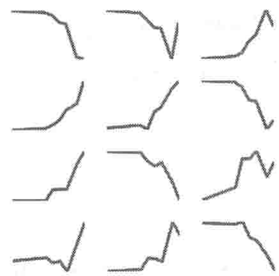
Another Example

(Taken from Kevin Murphy's ML textbook)

- Data: gene expression levels $\rightarrow [\;] \in \mathbb{R}^7$
- Goal: cluster genes with similar expression trajectories

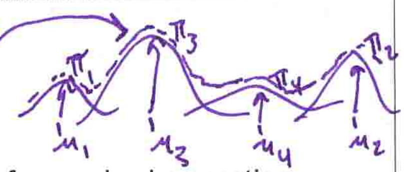


cluster means



Mixture models are useful for…

- Density estimation
  - Allows for multimodal density



- Clustering
  - Want membership information for each observation
    - e.g., topic of current document
  - Soft clustering: "responsibility of cluster k for point i"

$$p(z^i = k \mid x^i, \theta) = \frac{p(x^i \mid z^i = k, \phi)\, p(z^i = k \mid \pi)}{p(x^i \mid \theta)}$$

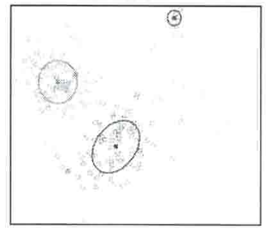
  - Hard clustering:

$$z^{i*} = \arg\max_k p(z^i = k \mid x^i, \theta) = \arg\max_k \log$$

$$= \arg\max_k \log p(x^i \mid z^i = k, \underset{\phi}{\theta}) + \log p(z^i = k \mid \underset{\pi}{\theta})$$

Issues

- Label switching
  - Color = label does not matter
  - Can switch labels and likelihood is unchanged



- Log likelihood is not convex in the parameters
  - No closed form gradient updates
  - Problem is simpler for "complete data likelihood"

IF we knew z^i

- More on this next time…

What you need to know

- Mixture model formulation
  - Generative model
  - Likelihood

ML Estimate of Mixture Model Params

- Log likelihood

$$L_x(\theta) \triangleq \log p(\{x^i\} \mid \theta) = \sum_i \log \sum_{z^i} p(x^i, z^i \mid \theta)$$

$p(x|\theta) = \prod_i p(x_i|\theta)$

- Want ML estimate

$$\hat{\theta}^{ML} = \arg\max_\theta L_x(\theta)$$

- Assume exponential family $p(x, z \mid \theta) = \frac{1}{Z(\theta)} e^{\theta' \phi(x,z)}$

$$L_x(\theta) = \sum_i \log \left(\sum_{z_i} e^{\theta^T \phi(z_i, x_i)} \right) - N \log Z(\theta)$$

- Neither convex nor concave and local optima

If "complete" data were observed…

- Assume class labels z^i were observed in addition to x^i

$$L_{x,z}(\theta) = \sum_i \log p(x^i, z^i \mid \theta) = \sum_i \log p(x_i|z_i,\theta) + \underbrace{\log p(z_i|\theta)}_{\pi_{z_i}}$$

$$= \sum_k \sum_{i:z_i=k} \log p(x_i|z_i,\phi_k) + \sum_{j=1}^{K-1} N_j \log \pi_j + N_K \log\left(1 - \sum_{j=1}^{K-1} \pi_j\right)$$

$|\{i : z_i = j\}|$ $\quad \sum \pi_j = 1$

- Compute ML estimates
  - Separates over clusters k!

$$\hat{\phi}_k = \arg\max_{\phi_k} \sum_{i:z_i=k} \log p(x_i|z_i,\phi_k) \qquad \hat{\pi}_k = \frac{N_k}{N}$$

- Example: mixture of Gaussians (MoG) $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i:z_i=k} x_i \qquad \hat{\pi}_k = \frac{N_k}{N}$$

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{i:z_i=k} x_i x_i^T - \hat{\mu}_k \hat{\mu}_k^T$$

Iterative Algorithm

- Motivates a coordinate ascent-like algorithm:
  1. Infer missing values z^i given estimate of parameters $\hat{\theta}$
  2. Optimize parameters to produce new $\hat{\theta}$ given "filled in" data z^i
  3. Repeat

$z_i \underset{max}{\overset{estimate}{\rightleftarrows}} \theta$

- Example: MoG (derivation soon… + HW)
  1. Infer "responsibilities"

"responsibility" of cluster k for point i

$$r_{ik} = p(z^i = k \mid x^i, \hat{\theta}^{(t-1)}) = \frac{\pi_k^{(t-1)} p(x_i \mid \phi_k^{(t-1)})}{\sum_j \pi_j^{(t-1)} p(x_i \mid \phi_j^{(t-1)})}$$

  2. Optimize parameters

$\max \text{ w.r.t. } \pi_k :$ $\quad \pi_k^{(t)} = \frac{1}{N}\sum_i r_{ik} = \frac{r_k}{N}$ ← soft counts!

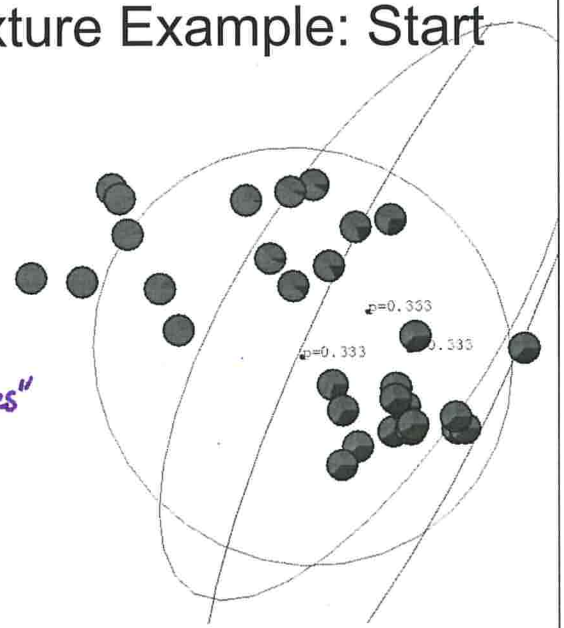
$\max \text{ w.r.t. } \phi_k :$

$\mu_k^{(t)} = \frac{\sum_i r_{ik} x_i}{r_k}$ ← weighted mean $\quad \Sigma_k^{(t)} = \frac{1}{r_k}\sum_i r_{ik} x_i x_i^T - \mu_k^{(t)} \mu_k^{(t)T}$

Gaussian Mixture Example: Start

Start with initial estimate of $\pi^{(0)}, \phi^{(0)}$

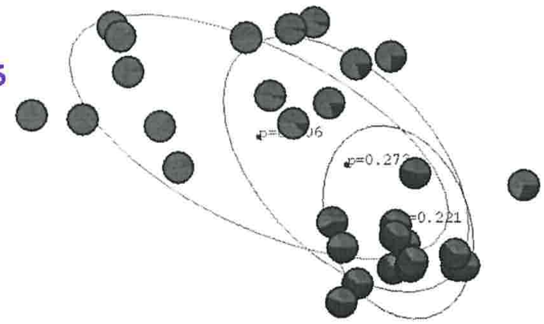
→ leads to initial "responsibilities"



After first iteration

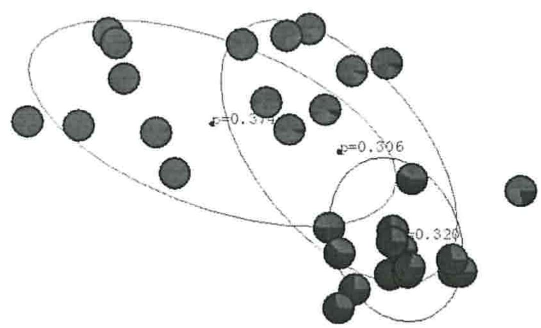
maximize
likelihood given
soft assignments

→ use new
$\pi^{(1)}, \phi^{(1)}$
to compute
new r_{ix}

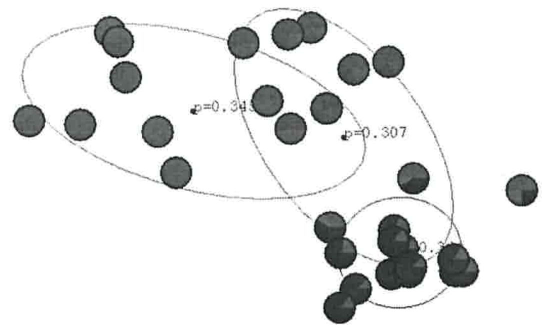


After 2nd iteration

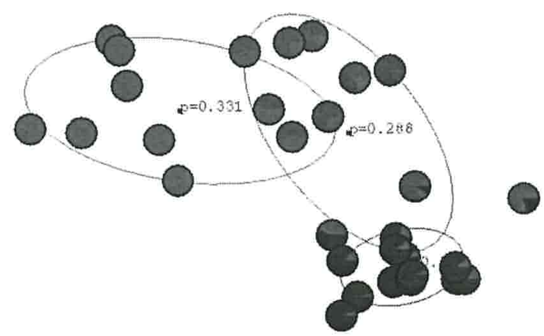
Iterate



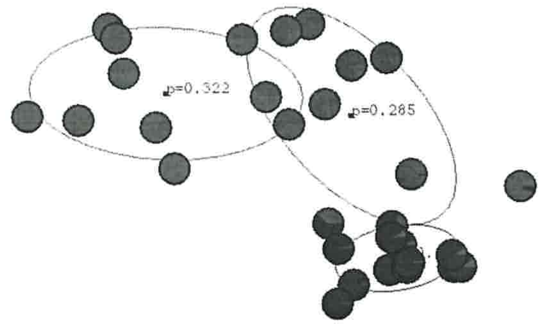
After 3rd iteration



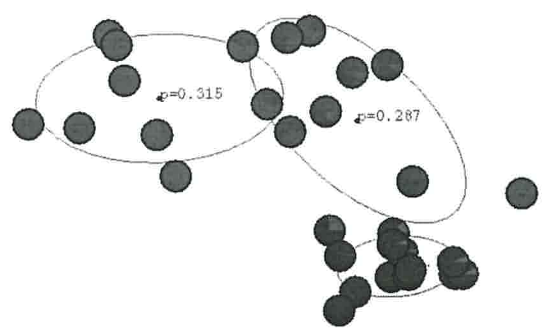
After 4th iteration



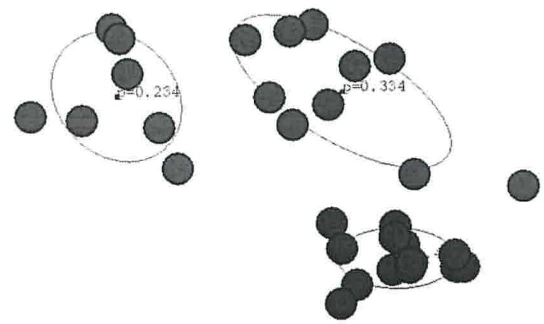
After 5th iteration



After 6th iteration

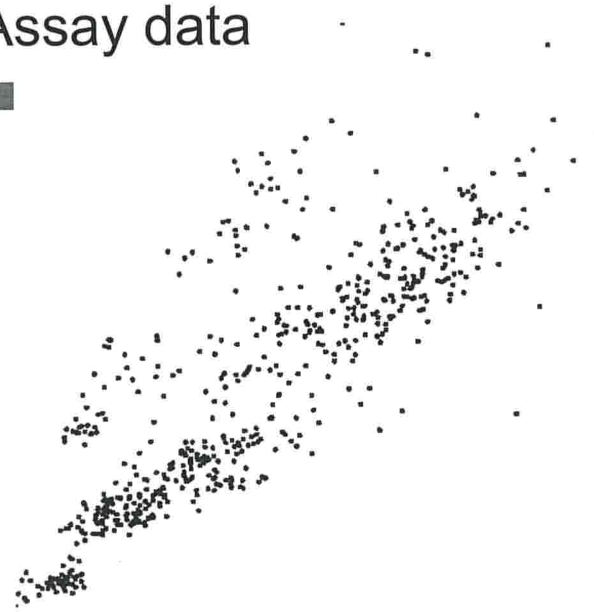


After 20th iteration



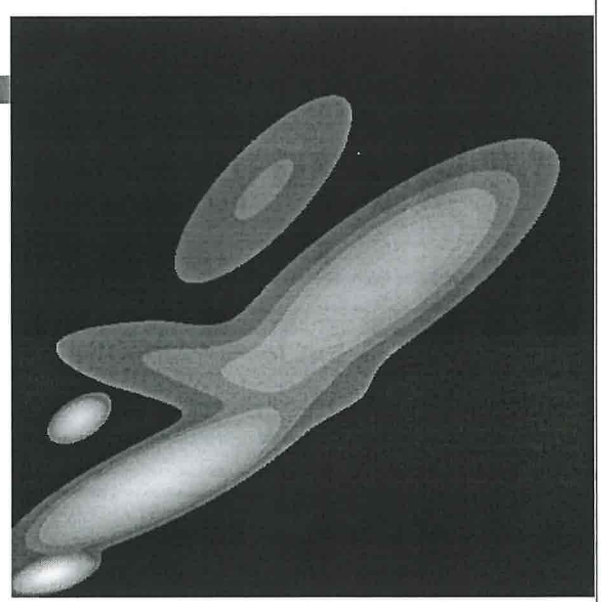
Looks pretty good!

Some Bio Assay data



GMM clustering of the assay data

Resulting Density Estimator



recall that GMMs can be used for density estimation

Expectation Maximization (EM) – Setup

- More broadly applicable than just to mixture models considered so far
- Model: x observable – *"incomplete" data* ← what we actually have
  y not (fully) observable – *"complete" data* ← what we wish we had
  θ parameters
- Interested in maximizing (wrt θ):

$$p(x \mid \theta) = \sum_y p(x, y \mid \theta)$$

- Special case: non-invertible deterministic fcn

$$x = g(y)$$

e.g. $y = \begin{bmatrix} z \\ x \end{bmatrix}$ ← class labels ← observations in standard mixture model

©Emily Fox 2013 27

Expectation Maximization (EM) – Derivation

- Step 1
  - Rewrite desired likelihood in terms of complete data terms

$$p(y \mid \theta) \underset{x=g(y)}{=} p(y \mid x, \theta) p(x \mid \theta)$$

$$\Rightarrow \underbrace{\log p(x \mid \theta)}_{L_x(\theta)} = \log p(y \mid \theta) - \log p(y \mid x, \theta)$$

- Step 2
  - Assume estimate of parameters $\hat{\theta}$
  - Take expectation with respect to $p(y \mid x, \hat{\theta})$ "$E[\cdot \mid x, \hat{\theta}]$"

$$L_x(\theta) = \underbrace{E[\log p(y \mid \theta) \mid x, \hat{\theta}]}_{U(\theta, \hat{\theta})} + \underbrace{E[-\log p(y \mid x, \theta) \mid x, \hat{\theta}]}_{V(\theta, \hat{\theta})}$$

©Emily Fox 2013 28

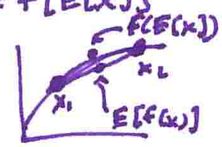
Expectation Maximization (EM) – Derivation

- Step 3
  - Consider log likelihood of data at any θ relative to log likelihood at $\hat{\theta}$

$$L_x(\theta) - L_x(\hat{\theta}) = [U(\theta,\hat{\theta}) - U(\hat{\theta},\hat{\theta})] + [V(\theta,\hat{\theta}) - V(\hat{\theta},\hat{\theta})]$$

- **Aside: Gibbs Inequality** $E_p[\log p(x)] \geq E_p[\log q(x)]$ $\forall q(\cdot)$

Proof: Use Jensen's Ineq $E[f(x)] \leq f[E[x]]$ for any concave $f(\cdot)$



Here:

$$E_p[\log q] - E_p[\log p] = E_p\left[\log \frac{q}{p}\right]$$

$$\leq \log E_p\left[\frac{q}{p}\right] = \log \int_x p(x) \frac{q(x)}{p(x)} dx = \log 1 = 0$$

Expectation Maximization (EM) – Derivation

$$L_x(\theta) - L_x(\hat{\theta}) = [U(\theta,\hat{\theta}) - U(\hat{\theta},\hat{\theta})] - \underbrace{[V(\theta,\hat{\theta}) - V(\hat{\theta},\hat{\theta})]}_{\geq 0}$$

- Step 4
  - Determine conditions under which log likelihood at θ exceeds that at $\hat{\theta}$

Using Gibbs inequality: $V(\theta,\hat{\theta}) = E[-\log p(y|x,\theta) | x, \hat{\theta}]$

$\geq E[-\log p(y|x,\hat{\theta}) | x, \hat{\theta}]$

$= V(\hat{\theta},\hat{\theta}) \quad \forall \theta$

If $U(\theta,\hat{\theta}) \geq U(\hat{\theta},\hat{\theta})$

Then

$$L_x(\theta) \geq L_x(\hat{\theta})$$

choosing a θ s.t. this is true means we're moving in the right direction (or not wrong!)

Motivates EM Algorithm

- Initial guess: $\hat{\theta}^{(0)}$
- Estimate at iteration *t*: $\hat{\theta}^{(t)}$

- **E-Step**

  Compute $U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}]$

- **M-Step**

  Compute $\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)})$

From before, $U(\hat{\theta}^{(t+1)}, \hat{\theta}^{(t)}) \geq U(\hat{\theta}^{(t)}, \hat{\theta}^{(t)})$

$\Rightarrow L_x(\hat{\theta}^{(t+1)}) \geq L_x(\hat{\theta}^{(t)})$

Example – Mixture Models

- **E-Step** Compute $U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}]$
- **M-Step** Compute $\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)})$

- Consider $y^i = \{z^i, x^i\}$ i.i.d.

$$p(x^i, z^i \mid \theta) = \pi_{z^i} p(x^i \mid \phi_{z^i}) = \prod_{k=1}^{K} (\pi_k p(x^i \mid \phi_k))^{I(z^i = k)}$$

$$E_{q_t}[\log p(y \mid \theta)] = \sum_i E_{q_t}[\log p(x^i, z^i \mid \theta)] = \cancel{\sum_i E_{q_t}\left[\sum_{k=1}^{K} \log \ldots\right]}$$

$$\overset{\shortparallel}{U(\theta, \hat{\theta}^{(t)})}$$

$$= \sum_i \sum_k E_{q_t}[I(z^i = k)] \log [\pi_k p(x^i \mid \phi_k)]$$

E-step compute these

$$= \sum_i \sum_k p(z_i = k \mid x_i, \hat{\theta}^{(t)}) \log [\pi_k p(x^i \mid \phi_k)]$$

M-step maximize

$$= \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} \log p(x_i \mid \phi_k)$$

Coordinate Ascent Behavior

- Bound log likelihood:

$$L_x(\theta) = U(\theta,\hat\theta^{(t)}) + V(\theta,\hat\theta^{(t)})$$
$$\geq U(\theta,\hat\theta^{(t)}) + V(\hat\theta^{(t)},\hat\theta^{(t)}) \triangleq LB_x(\theta,\hat\theta^{(t)})$$
$$L_x(\hat\theta^{(t)}) = U(\hat\theta^{(t)},\hat\theta^{(t)}) + V(\hat\theta^{(t)},\hat\theta^{(t)}) = LB_x(\hat\theta^{(t)},\hat\theta^{(t)})$$

tight at $\hat\theta^{(t)}$

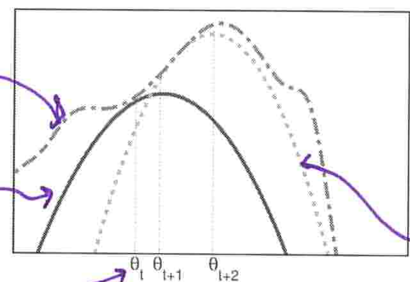


Figure from KM textbook

Comments on EM

- Since Gibbs inequality is satisfied with equality only if *p=q*, any step that changes θ should strictly **increase likelihood**
  assuming identifiability (i.e. $\nexists\, \theta \neq \theta'$ s.t. $p(x|\theta) = p(x|\theta')$)
- In practice, can replace the **M-Step** with increasing *U* instead of maximizing it (**Generalized EM**)
  exact max can be hard to compute
- Under certain conditions (e.g., in exponential family), can show that EM **converges to a stationary point** of $L_x(\theta)$
- Often there is a **natural choice for *y*** … has physical meaning
  like in mix model with $y = \{z, x\}$ "cluster assign."
- If you want to choose any *y*, not necessarily *x=g(y)*, replace $p(y \mid \theta)$ in *U* with $p(y, x \mid \theta)$

Initialization

- In mixture model case where $y^i = \{z^i, x^i\}$ there are many ways to initialize the EM algorithm
- Examples:
  - Choose K observations at random to define each cluster. Assign other observations to the nearest "centriod" to form initial parameter estimates
  - Pick the centers sequentially to provide good coverage of data
  - Grow mixture model by splitting (and sometimes removing) clusters until K clusters are formed

  ... many choices!
- Can be quite important to convergence rates in practice

  and quality of local optima found

MAP Estimation

- Bayesian approach:
  - Place **prior** $p(\theta)$ on parameters
  - Infer **posterior** $p(\theta \mid x)$ $= \dfrac{p(x|\theta)p(\theta)}{p(x)}$
- Many, many, many motivations and implications
  - For the sake of this ~~call~~ class, simplest motivation is to think of this as akin to regularization

$$\hat{\theta}^{MAP} = \arg\max_\theta \log p(\theta \mid x) = \arg\max_\theta \underbrace{\log p(x|\theta)}_{\text{ML term}} + \log p(\theta)$$

reg.

  - Saw importance of regularization in logistic regression (ML estimate can overfit data and lead to poor generalization)

EM Algorithm – MAP Case

- Re-derive EM algorithm for $p(\theta \mid x)$
- Add $\log p(\theta)$ to $U(\theta, \hat{\theta}^{(t)})$

Prev $E[\log p(y|\theta) | x, \hat{\theta}]$
Now $E[\log p(y|\theta) | x, \hat{\theta}] + \log p(\theta)$

  - What must be computed in E-Step remains unchanged because this term does not depend on *y*.
  - M-Step becomes:

$$\hat{\theta}^{(t+1)} = \arg\max_{\theta} U(\theta, \hat{\theta}^{(t)}) + \log p(\theta)$$

affects max w.r.t. θ

MAP EM Example – MoG

$p(\theta|y)$ in same family as $p(\theta)$

- For mixture of Gaussians, conjugate priors are:

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K) \quad \{\mu_k, \Sigma_k\} \sim \mathrm{NIW}(m_0, \kappa_0, \nu_0, S_0)$$

- Results in following M-Step:

$\hat{\mu}_k$ from before
mean of pseudo-obs.
pseudocounts of obs in cluster k

$$\hat{\mu}_k = \frac{r_k \bar{x}_k + \kappa_0 m_0}{r_k + \kappa_0} \qquad \hat{\pi}_k = \frac{r_k + \alpha_k - 1}{N + \sum_k \alpha_k - K}$$

$\hat{\Sigma}_k$ from before

$$\hat{\Sigma}_k = \frac{S_0 + r_k S_k + \frac{\kappa_0 r_k}{\kappa_0 + r_k}(\bar{x}_k - m_0)(\bar{x}_k - m_0)'}{\nu_0 + r_k + d + 2}$$

dimension

What you need to know

- Mixture model formulation
  - Generative model
  - Likelihood
- Expectation Maximization (EM) Algorithm
  - Derivation
  - Concept of non-decreasing log likelihood
  - Application to standard mixture models

Course Announcements

- Homework 2 will be posted on Thursday
  - Due 2 weeks later (Feb 14)
- Project proposals:
  - Initial ideas now posted
  - Deadline extended to Tues, Feb 5
  - 1 page, 1-2 people
- Recitation on Thursday (Linda)
- Office hours as normal