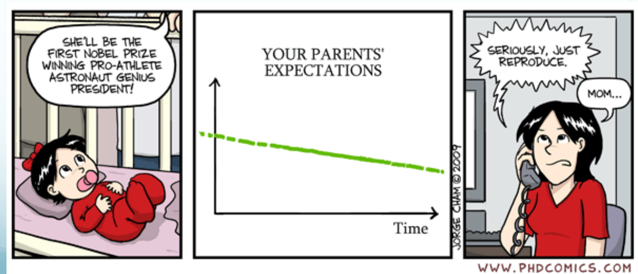
The EM Algorithm

Mark Dredze

Machine Learning
CS 600.475



Similarities

	K-Means	Gaussian Mixtures
Assign examples to clusters	r_{nk}	$\gamma(z_k)$
Compute new model parameters that maximize assignments	μ_k	$\mu_k \quad \Sigma_k \quad \pi_k$

Same Algorithm

- The maximization algorithm for both models is the same!
- Iterate two steps
  - Compute the **expected** cluster assignments according to the current model
  - **Maximize** the model parameters according to the current cluster assignments
- Expectation Maximization Algorithm (EM)

EM Algorithm

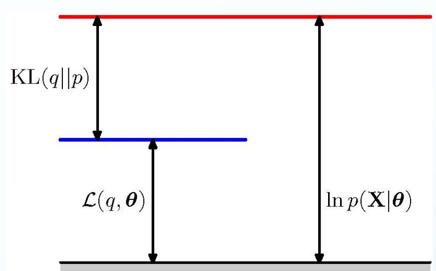
- A general technique for maximizing likelihood when you have latent variables
  - Latent variables: a variable you do not observe
  - We never get to see examples of cluster assignments
- EM allows us to write objectives without seeing these variables
  - Maximization step is familiar
    - Find the best parameters given the observations
  - Expectation step is new!
    - Pretend we see the latent variables

EM Algorithm + Clustering

- Clustering is a great example of an EM algorithm
- We could easily maximize the objective if we only knew the hidden variables
- Compute the **expected** cluster assignments, then update
- Not just clustering!
  - EM is a very general algorithm used all over

General EM Algorithm

Pictorial View

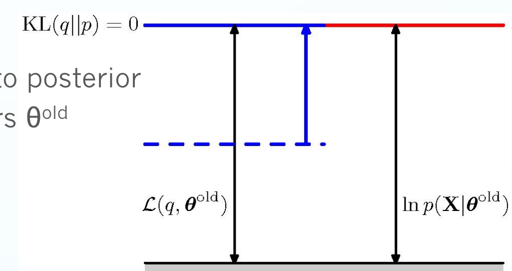


- We can decompose the likelihood as

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

Pictorial View

- E-Step:
  - q distribution set to posterior of current parameters θ^{old}



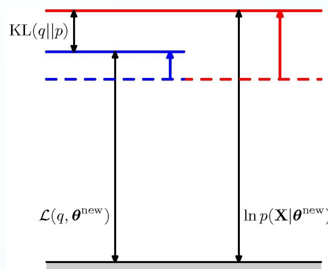
$$\log p(X \mid \theta) = L(q,\theta) + \underbrace{\boxed{KL(q \,||\, p)}}_{\text{Goes to 0}}$$

$$\log p(X \mid \theta) = L(q,\theta) + 0$$

The “observed” variables Z match what we **expect** given parameters

Pictorial View

- M-Step:
  - Maximize L(q,θ) by finding new θ for fixed q(Z)



$$\log p(X \mid \theta) = \underbrace{\boxed{L(q,\theta)}}_{\text{Increases}} + \underbrace{\boxed{KL(q \,\|\, p)}}_{\text{Can only increase}}$$

$$\log p(X \mid \theta) \geq \log p(X \mid \theta^{old})$$

The new parameters θ best explain the "observed" variables Z

Convergence

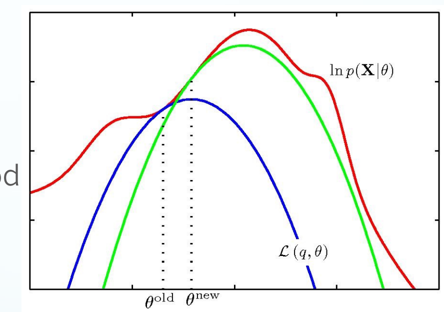
- When will $\log p(X \mid \theta) = \log p(X \mid \theta^{old})$?
  - When we can no longer increase the likelihood
  - Since we likelihood always increases, this must be a maximum (possibly local)

Convergence

- We now see why EM converges in general
  - We are always increasing the likelihood function
  - At some point we won't be able to increase it any more
- Very powerful result
  - For any problem with latent variables, if you can write the complete data likelihood, you can use EM
  - The algorithm will always converge!

Pictorial View

- The likelihood function (red)
- Using old parameters lower bound the likelihood using L (blue)
- Maximize L to get new parameters



- Next E step gives new lower bound (green)

Examining EM

The General EM Algorithm

- Goal: maximize a likelihood function $p(X|\theta)$
  - Write a joint distribution over the complete data $p(X,Z|\theta)$
- Choose an initial setting for θ^{old}
- **E step** Compute the $q(Z)$ as $p(Z|X, \theta^{old})$
- **M step** Compute θ^{new} given by $\theta^{new} = \arg\max_\theta Q(\theta, \theta^{old})$

$$Q(\theta, \theta^{old}) = \sum_Z p(Z|X, \theta^{old}) \log p(X, Z|\theta)$$

- Let $\theta^{old} = \theta^{new}$
- Repeat until convergence

GMMs with EM

EM is Everywhere

- Remember the similar forms of GMM and K-means?
  - K-means is an application of EM in the limit
  - Force hard cluster assignments
- See, EM really is everywhere
  - Google scholar: Dempster, et al. Maximum likelihood from incomplete data via the EM algorithm.
    - 22083 citations

General EM

- The EM form is the same, but each step can be more complicated
- E step
  - Finding the values for the hidden variables may not be easy
    - We may need to approximate the values
- M step
  - Maximization may require multiple steps, optional constraints

Latent Variables

- EM is useful for latent variables
  - Variables that you do not observe
- What is the structure of these latent variables?
  - How do they influence the observed variables?
  - Can you have multiple latent variables in a complex structure?
- We need some way to talk about these variables formally

Next Time

Graphical Models