STAT 215A Fall 2020 Week 8

James Duncan, OH: M, Th 2-4pm

Announcements

- Lab 3 released this morning
  - DUE: **10/22 at 11:59pm** (only 10 days!)
- Midterm: 10/29
  - Will release practice midterm this Friday & go over solutions in lab on 10/23
  - Review lecture Thursday 10/22

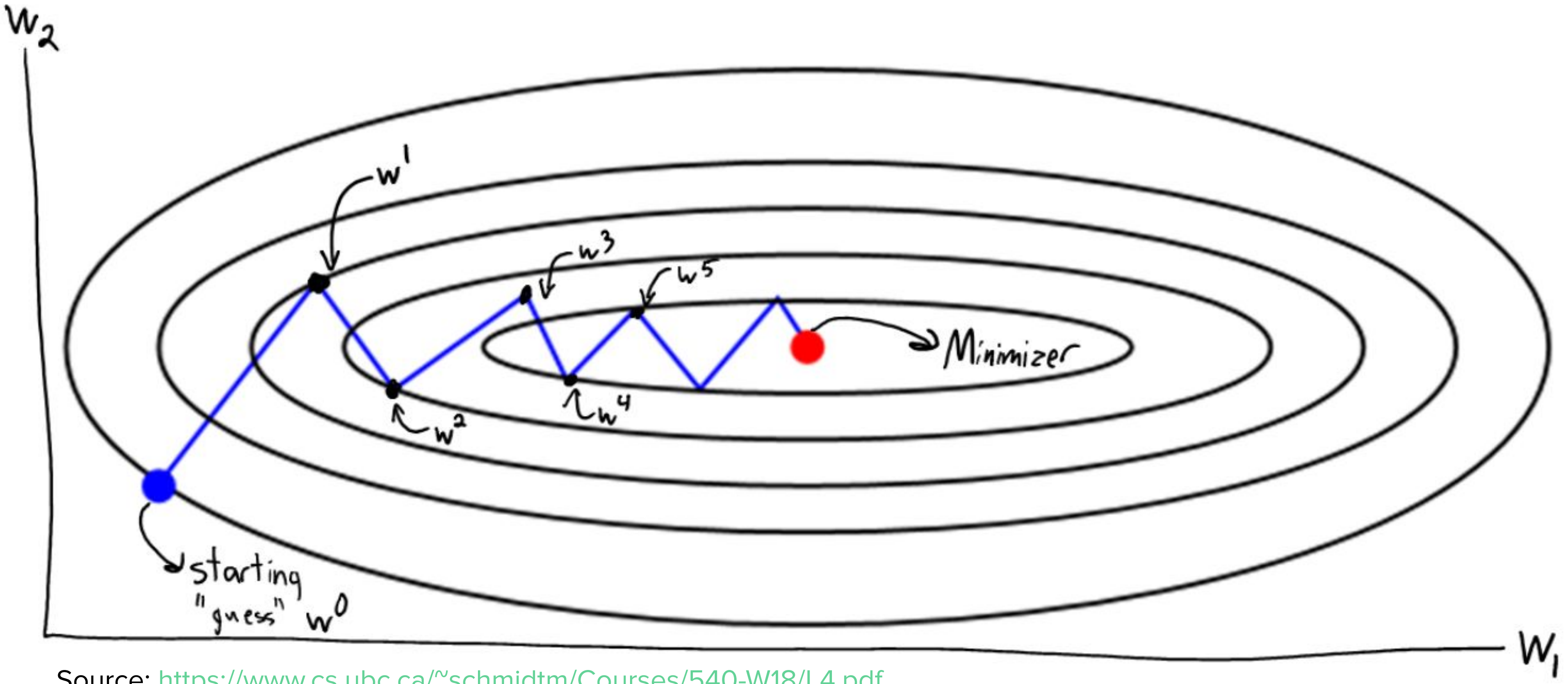
Lab 3: Stability of K-means + Computability

- Sign up for an SCF account if you haven't yet (required)
- Don't wait until the last minute
  - SCF could be busy and code will take time to run
- No need to do PCA; just apply k-means to raw `lingBinary` data
- You can do better than the Figure 3 in Ben-Hur
- Either manually copy over the `data/` folder to SCF or push it to GitHub and then remove it
- While the writeup is shorter than usual, there will still be a writing component of the grade
- Take a look at Google R Style Guide and Part I Analyses Section of the Tidyverse Style Guide
- If you are using the SCF JupyterHub, be sure to "Stop Server" when you are done

Outline for today

- EM algorithm

Gradient descent (in 2D)



Source: https://www.cs.ubc.ca/~schmidtm/Courses/540-W18/L4.pdf

EM Algorithm

Dempster et al., 1977

Various resources

Intuition:

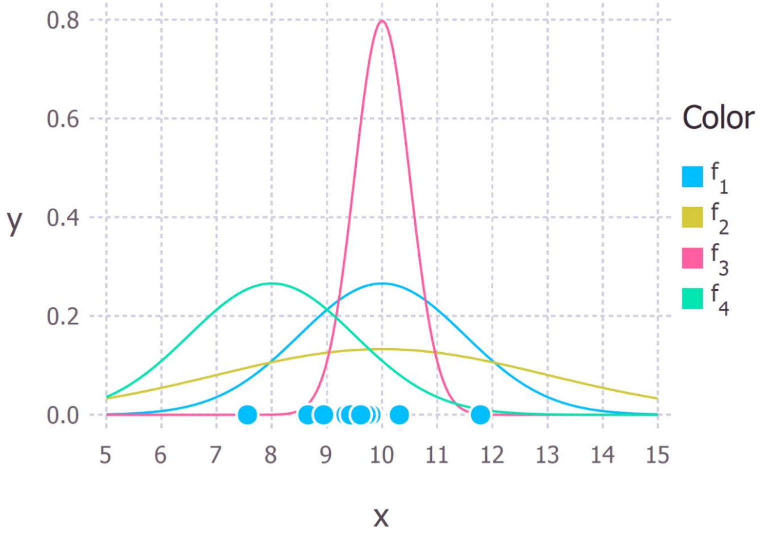
- Quick: https://stackoverflow.com/questions/11808074/what-is-an-intuitive-explanation-of-the-expectation-maximization-technique#answer-43561339

Math:

- Quick overview: http://www.seanborman.com/publications/EM_algorithm.pdf
  - Included in the `week8` folder
- More in-depth: https://arxiv.org/pdf/1105.1476.pdf

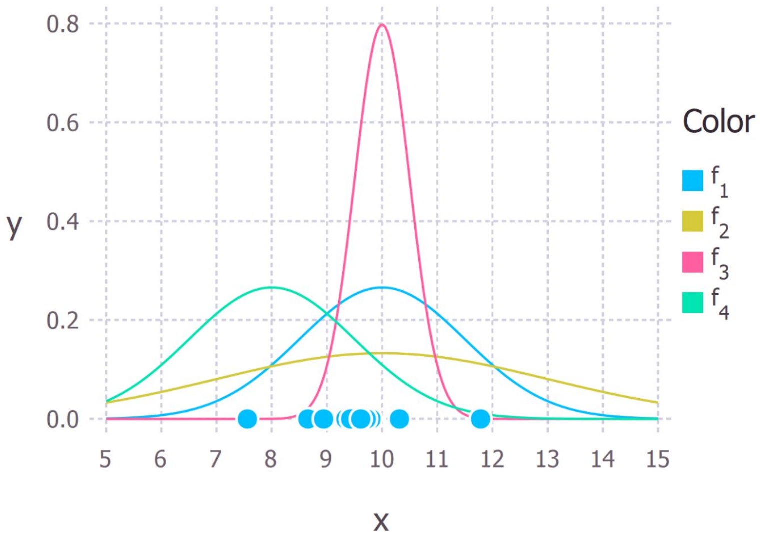
Maximum Likelihood Estimation Review

- Data: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_\theta(x)$
-



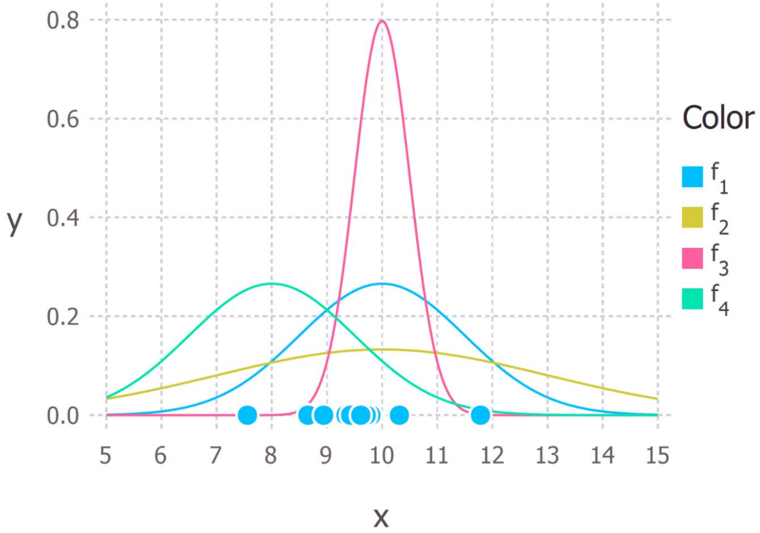
Maximum Likelihood Estimation Review

- Data: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_\theta(x)$
- Likelihood: $\mathcal{L}(\theta) = \prod_{i=1}^{n} p_\theta(x_i)$
- Log-likelihood: $\ell(\theta) = \log \mathcal{L}(\theta)$
-



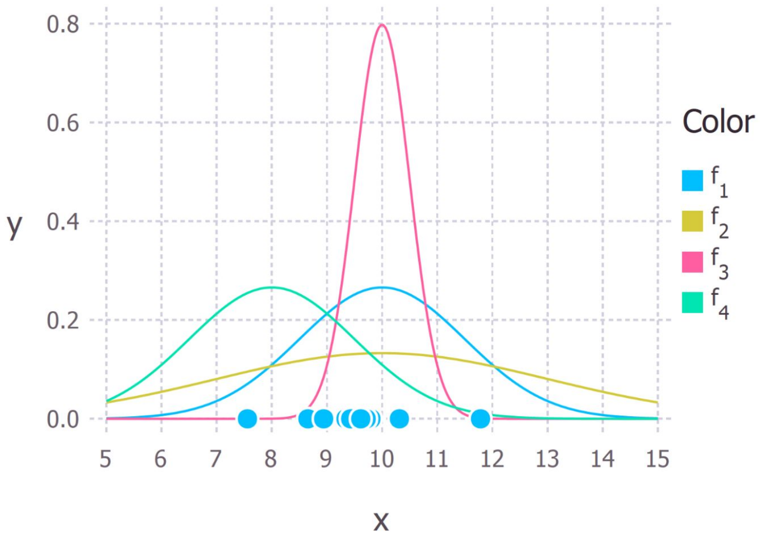
Maximum Likelihood Estimation Review

- Data: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_\theta(x)$
- Likelihood: $\mathcal{L}(\theta) = \prod_{i=1}^{n} p_\theta(x_i)$
- Log-likelihood: $\ell(\theta) = \log \mathcal{L}(\theta)$
- MLE: $$\begin{aligned}\hat{\theta} &= \arg\max_\theta \mathcal{L}(\theta) \\ &= \arg\max_\theta \ell(\theta)\end{aligned}$$
-



Maximum Likelihood Estimation Review

- Data: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_\theta(x)$
- Likelihood: $\mathcal{L}(\theta) = \prod_{i=1}^{n} p_\theta(x_i)$
- Log-likelihood: $\ell(\theta) = \log \mathcal{L}(\theta)$
- MLE: $$\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta) = \arg\max_\theta \ell(\theta)$$



- **Intuition**: Find the value of θ under which we would be least surprised to see a sample like the observed one.
- **Optimization problem**: take derivative, set equal to zero, solve for parameter.

EM Overview

Motivation 1: “Hard” Maximum Likelihood Estimation Problems

Say we have the following mixture of Gaussians problem:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2),\ \pi_1 + \pi_2 = 1$$

EM Overview

Motivation 1: “Hard” Maximum Likelihood Estimation Problems

Say we have the following mixture of Gaussians problem:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2),\ \pi_1 + \pi_2 = 1$$

- Because of the sum of normals, the log-likelihood isn’t as helpful as before and taking derivatives w.r.t $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and setting equal to zero, etc., won’t be particularly successful.
-

EM Overview

Motivation 1: “Hard” Maximum Likelihood Estimation Problems

Say we have the following mixture of Gaussians problem:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2),\ \pi_1 + \pi_2 = 1$$

- Because of the sum of normals, the log-likelihood isn’t as helpful as before and taking derivatives w.r.t $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and setting equal to zero, etc., won’t be particularly successful.
- Instead, we can introduce **latent variables** which tell us which of the two Gaussians each observation comes from: $Z_i \overset{i.i.d.}{\sim} \text{Bernoulli}(1 - \pi_1) + 1$
-

EM Overview

Motivation 1: “Hard” Maximum Likelihood Estimation Problems

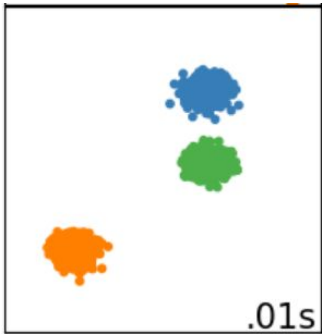
Say we have the following mixture of Gaussians problem:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2),\ \pi_1 + \pi_2 = 1$$

- Because of the sum of normals, the log-likelihood isn’t as helpful as before and taking derivatives w.r.t $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and setting equal to zero, etc., won’t be particularly successful.
- Instead, we can introduce **latent variables** which tell us which of the two Gaussians each observation comes from: $Z_i \overset{i.i.d.}{\sim} \text{Bernoulli}(1 - \pi_1) + 1$
- If we know the latent allocations then the problem simply becomes two easy MLE exercises.

$$\boxed{\begin{aligned} X_i | Z_i = 1 &\overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2) \\ X_i | Z_i = 2 &\overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2) \end{aligned}}$$

EM Overview



Motivation 2: Clustering

Since EM helps with finding solutions to mixture problems, it lends itself naturally to clustering.

Recall:

- K-means performs well when clusters are homogeneous.
-

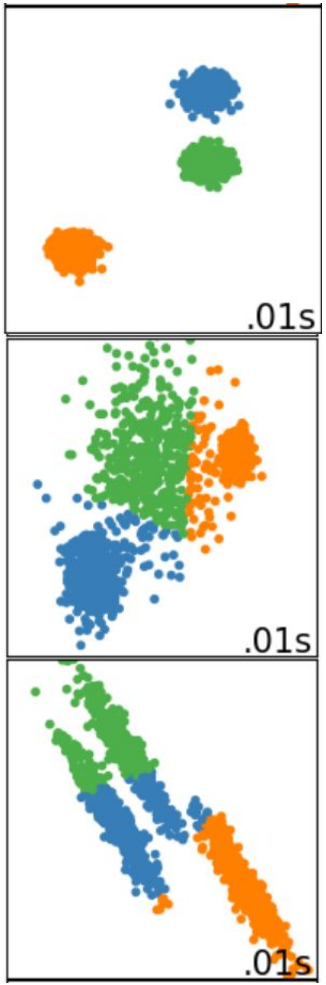
EM Overview

Motivation 2: Clustering

Since EM helps with finding solutions to mixture problems, it lends itself naturally to clustering.

Recall:

- K-means performs well when clusters are homogeneous.
- But if fails miserably when faced with anisotropy.



EM Overview

Motivation 2: Clustering

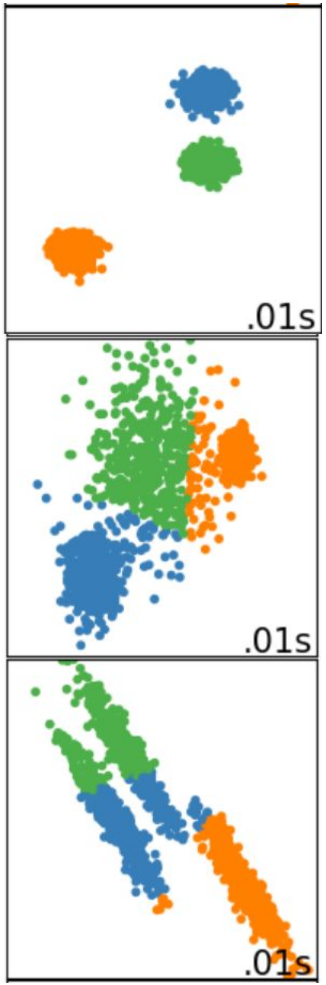
Since EM helps with finding solutions to mixture problems, it lends itself naturally to clustering.

Recall:

- K-means performs well when clusters are homogeneous.
- But if fails miserably when faced with anisotropy.

The EM generalizes K-means:

- Still performs great where K-means does.



EM Overview

Motivation 2: Clustering

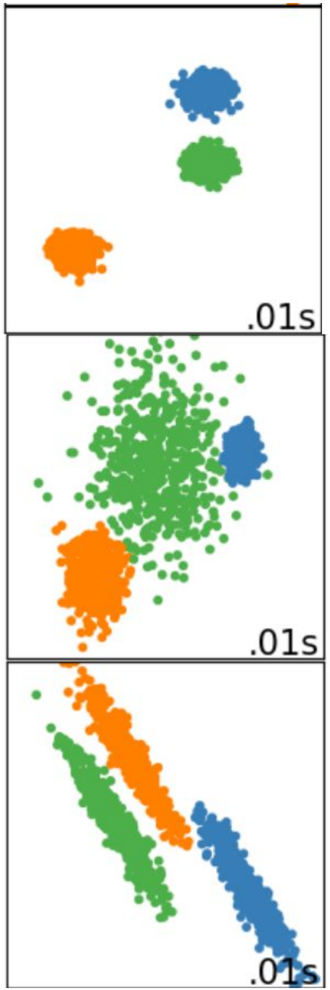
Since EM helps with finding solutions to mixture problems, it lends itself naturally to clustering.

Recall:

- K-means performs well when clusters are homogeneous.
- But if fails miserably when faced with anisotropy.

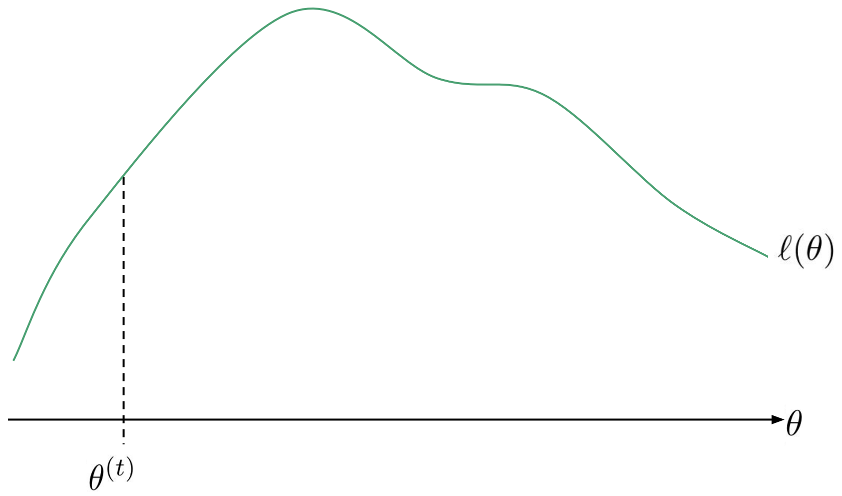
The EM generalizes K-means:

- Still performs great where K-means does.
- But it can also handle differences in spread and symmetry.
- EM is a “soft” clustering algorithm.



EM algorithm intuition

Say we make a guess $\theta^{(t)}$

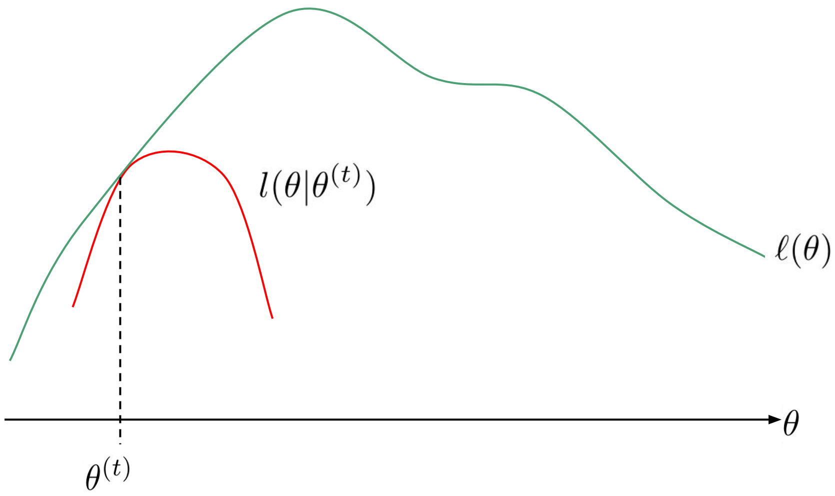


EM algorithm intuition

Say we make a guess $\theta^{(t)}$

The insight of the EM algorithm is that we can find a function $l(\theta|\theta^{(t)})$ such that

- $l(\theta|\theta^{(t)}) \leq \ell(\theta)$
- $l(\theta^{(t)}|\theta^{(t)}) = \ell(\theta^{(t)})$



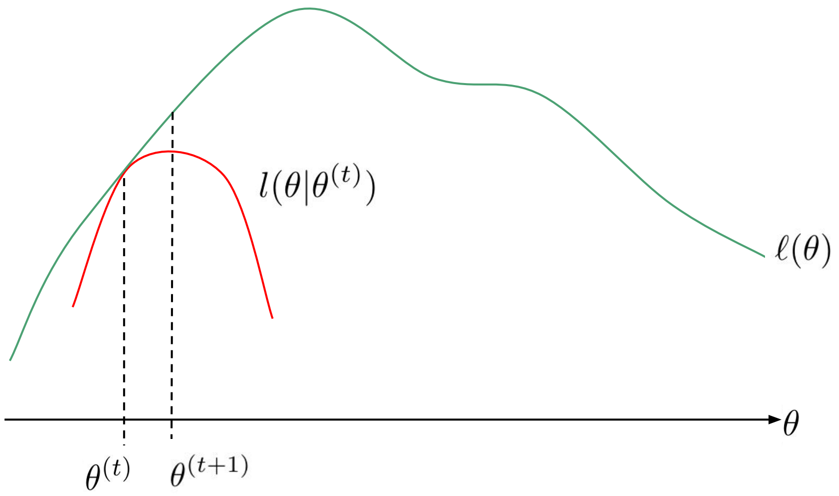
EM algorithm intuition

Say we make a guess $\theta^{(t)}$

The insight of the EM algorithm is that we can find a function $l(\theta|\theta^{(t)})$ such that

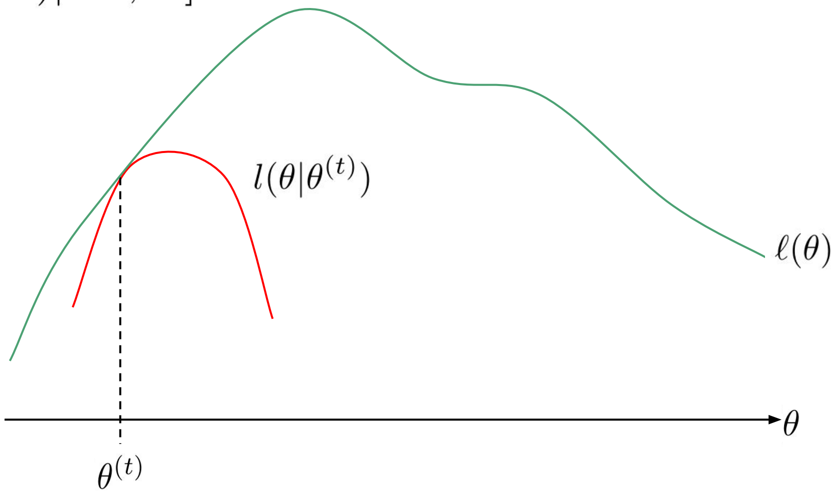
- $l(\theta|\theta^{(t)}) \leq \ell(\theta)$
- $l(\theta^{(t)}|\theta^{(t)}) = \ell(\theta^{(t)})$

So any θ that increases $l(\theta|\theta^{(t)})$ also increases $\ell(\theta)$.



EM algorithm steps

It turns out that maximizing $l(\theta|\theta^{(t)})$ is equivalent* to maximizing the expectation $Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; X, Z)|\theta^{(t)}, X]$



* see the Borman tutorial for the derivation

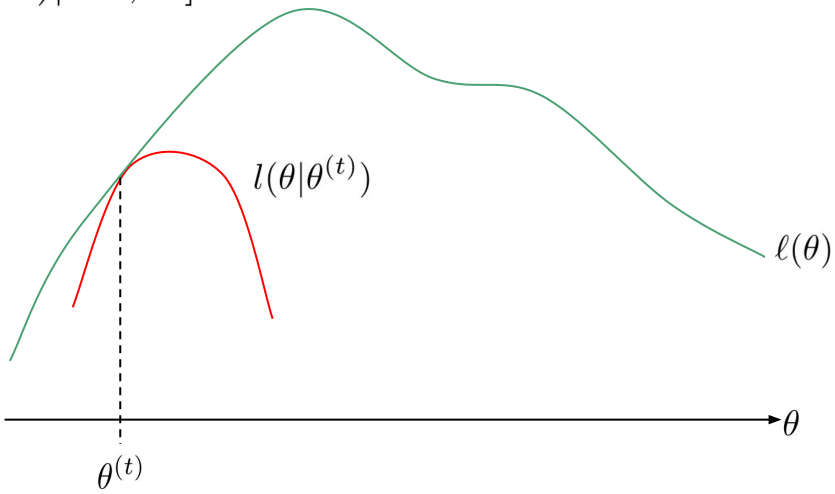
EM algorithm steps

It turns out that maximizing $l(\theta|\theta^{(t)})$ is equivalent* to maximizing the expectation $Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; X, Z)|\theta^{(t)}, X]$

The EM algorithm involves two steps that are repeated until convergence:

1. **E:** Calculate the expectation

$$Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; X, Z)|\theta^{(t)}, X]$$



* see the Borman tutorial for the derivation

EM algorithm steps

It turns out that maximizing $l(\theta|\theta^{(t)})$ is equivalent* to maximizing the expectation $Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; X, Z)|\theta^{(t)}, X]$

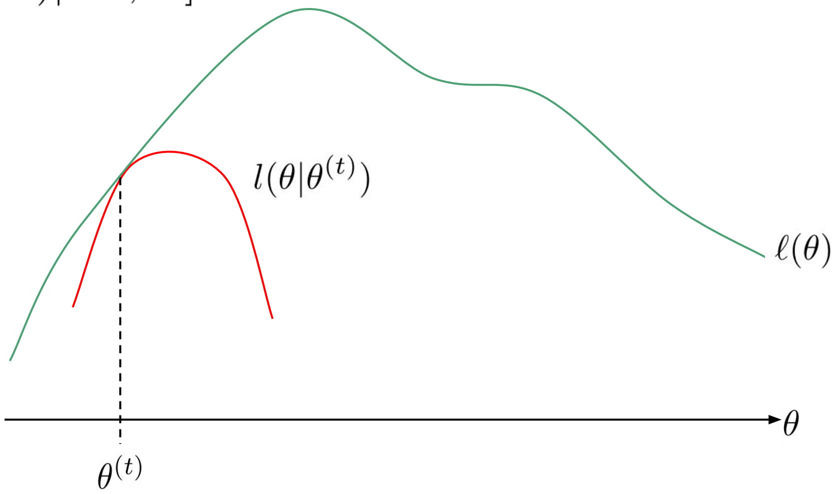
The EM algorithm involves two steps that are repeated until convergence:

1. **E:** Calculate the expectation

$$Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; X, Z)|\theta^{(t)}, X]$$

2. **M:** Maximize $Q(\theta|\theta^{(t)})$ w.r.t. θ

Note: we can initialize with a random guess $\theta^{(0)}$



* see the Borman tutorial for the derivation

EM: Gaussian Mixture Example

Same setup from before:

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2), p_1 + p_2 = 1$$

$$X_i | Z_i = 1 \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2) \qquad X_i | Z_i = 2 \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$$

$$Z_i \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi) + 1$$

EM: Gaussian Mixture Example

Same setup from before:

$$X_1, \dots, X_n \overset{i.i.d.}{\sim} \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2), p_1 + p_2 = 1$$

$$X_i | Z_i = 1 \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2) \qquad X_i | Z_i = 2 \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$$

$$Z_i \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi) + 1$$

Likelihood: $p_\theta(x_i, z_i) = p_\theta(x_i | z_i) p_\theta(z_i) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}\right\} \pi_1, & \text{if } Z_i = 1 \\ \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{(x_i - \mu_2)^2}{2\sigma_2^2}\right\} \pi_2, & \text{if } Z_i = 2 \end{cases}$

$$\log p_\theta(x_i, z_i) = \begin{cases} -\frac{1}{2} \log 2\pi - \log \sigma_1 - \frac{(x_i - \mu_1)^2}{2\sigma_1^2} + \log \pi_1, & \text{if } Z_i = 1 \\ -\frac{1}{2} \log 2\pi - \log \sigma_2 - \frac{(x_i - \mu_2)^2}{2\sigma_2^2} + \log \pi_2, & \text{if } Z_i = 2 \end{cases}$$

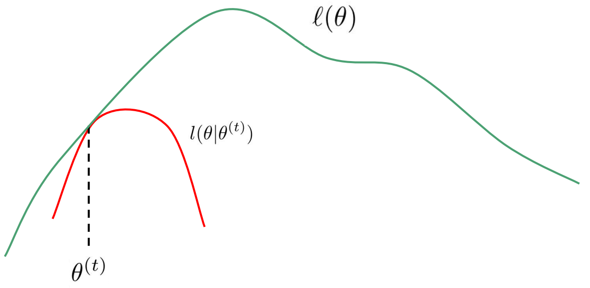
E-Step: Compute $Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; X, Z)|\theta^{(t)}, X]$

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \mathbb{E}[\ell(\theta; X_i, Z_i)|\theta^{(t)}, X]$$

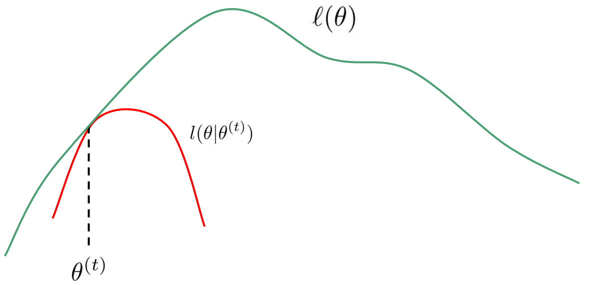
$$= \sum_{i=1}^{n} \Big\{ \mathbb{E}[\ell(\theta; X_i, Z_i)|\theta^{(t)}, X, Z_i = 1]\mathbb{P}(Z_i = 1|\theta^{(t)}, X) +$$

$$\mathbb{E}[\ell(\theta; X_i, Z_i)|\theta^{(t)}, X, Z_i = 2]\mathbb{P}(Z_i = 2|\theta^{(t)}, X) \Big\}$$

$$\text{(law of total expectation)}$$



E-Step: Compute $Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta; X, Z)|\theta^{(t)}, X]$



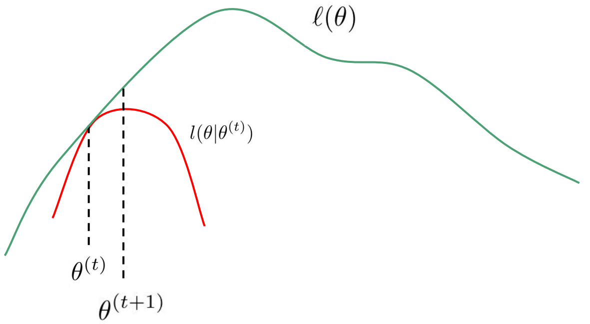
$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= \sum_{i=1}^{n} \mathbb{E}[\ell(\theta; X_i, Z_i)|\theta^{(t)}, X] \\
&= \sum_{i=1}^{n} \Big\{ \mathbb{E}[\ell(\theta; X_i, Z_i)|\theta^{(t)}, X, Z_i = 1]\mathbb{P}(Z_i = 1|\theta^{(t)}, X) + \\
&\qquad \mathbb{E}[\ell(\theta; X_i, Z_i)|\theta^{(t)}, X, Z_i = 2]\mathbb{P}(Z_i = 2|\theta^{(t)}, X) \Big\} \\
&\qquad \text{(law of total expectation)} \\
&= \sum_{i=1}^{n} \Big\{ \log \pi_1 - \frac{1}{2}\log 2\pi - \log \sigma_1 - \frac{(x_i - \mu_1)^2}{2\sigma_2^2} Z_{i,1}^{(t)} + \\
&\qquad \log \pi_2 - \frac{1}{2}\log 2\pi - \log \sigma_2 - \frac{(x_i - \mu_2)^2}{2\sigma_2^2} Z_{i,2}^{(t)} \Big\}
\end{aligned}
$$

$$
\boxed{
\begin{aligned}
Z_{i,j}^{(t)} &= \mathbb{P}(Z_i = j|\theta^{(t)}, X) \\
&= \frac{\pi_j^{(t)} \phi(\frac{x_i - \mu_j}{\sigma_j})/\sigma_j}{\sum_{k=1}^{2} \pi_k^{(t)} \phi(\frac{x_i - \mu_k}{\sigma_k})/\sigma_k)/\sigma_k}
\end{aligned}
}
$$

Standard normal pdf

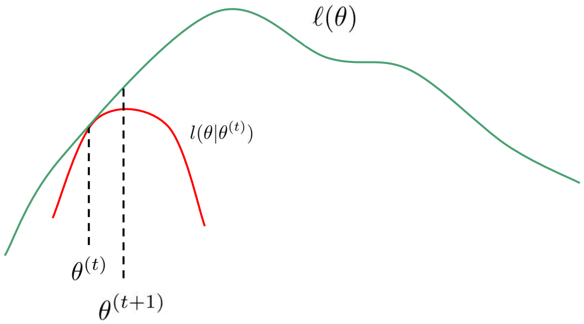
M-Step: Maximize $Q(\theta|\theta^{(t)})$ w.r.t. $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$

- π_1: $\dfrac{\partial Q}{\partial \pi_1} = \dfrac{\partial}{\partial \pi_1} \sum_{i=1}^{n} \left(\log \pi_1 Z_{i,1}^{(t)} + \log(1-\pi_1) Z_{i,2}^{(t)} \right)$



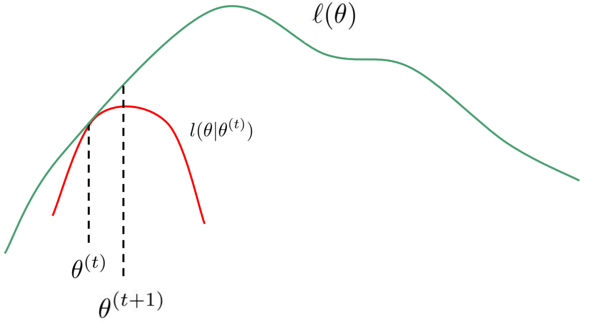
M-Step: Maximize $Q(\theta|\theta^{(t)})$ w.r.t. $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$

- $\pi_1:$ $$\frac{\partial Q}{\partial \pi_1} = \frac{\partial}{\partial \pi_1} \sum_{i=1}^{n} \left(\log \pi_1 Z_{i,1}^{(t)} + \log(1-\pi_1) Z_{i,2}^{(t)} \right)$$
$$= \sum_{i=1}^{n} \frac{Z_{i,1}^{(t)}}{\pi_1} + \sum_{i=1}^{n} \frac{Z_{i,2}^{(t)}}{1-\pi_1}$$



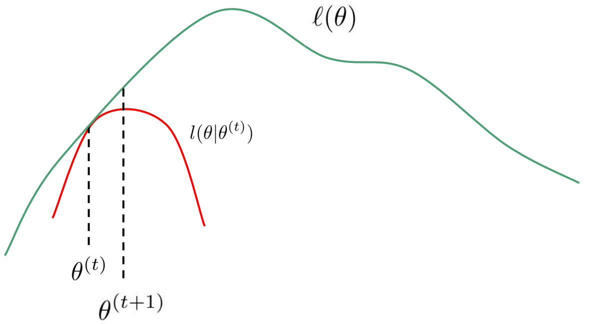
M-Step: Maximize $Q(\theta|\theta^{(t)})$ w.r.t. $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$

- π_1: $$\frac{\partial Q}{\partial \pi_1} = \frac{\partial}{\partial \pi_1} \sum_{i=1}^{n} \left(\log \pi_1 Z_{i,1}^{(t)} + \log(1-\pi_1) Z_{i,2}^{(t)} \right)$$
$$= \sum_{i=1}^{n} \frac{Z_{i,1}^{(t)}}{\pi_1} + \sum_{i=1}^{n} \frac{Z_{i,2}^{(t)}}{1-\pi_1}$$
$$\overset{set}{=} 0 \implies \boxed{\pi_1^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)}}{\sum_i Z_{i,1}^{(t)} + Z_{i,2}^{(t)}} = \frac{1}{n} \sum_i Z_{i,1}^{(t)}}$$



M-Step: Maximize $Q(\theta|\theta^{(t)})$ w.r.t. $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$

- $\pi_1:$
$$\frac{\partial Q}{\partial \pi_1} = \frac{\partial}{\partial \pi_1} \sum_{i=1}^{n} \left(\log \pi_1 Z_{i,1}^{(t)} + \log(1-\pi_1) Z_{i,2}^{(t)} \right)$$
$$= \sum_{i=1}^{n} \frac{Z_{i,1}^{(t)}}{\pi_1} + \sum_{i=1}^{n} \frac{Z_{i,2}^{(t)}}{1-\pi_1}$$
$$\overset{set}{=} 0 \implies \boxed{\pi_1^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)}}{\sum_i Z_{i,1}^{(t)} + Z_{i,2}^{(t)}} = \frac{1}{n} \sum_i Z_{i,1}^{(t)}}$$



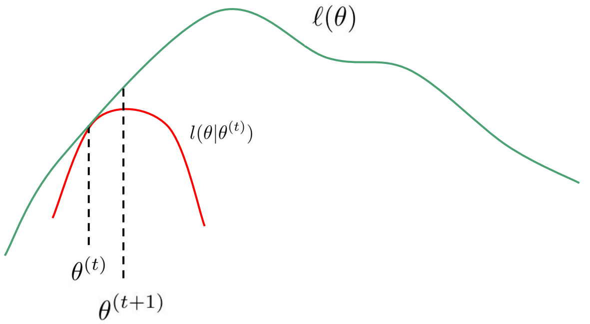
- $\mu_1:$
$$\frac{\partial Q}{\partial \mu_1} = \frac{\partial}{\partial \mu_1} \sum_{i=1}^{n} \left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) Z_{i,1}^{(t)} \implies \boxed{\mu_1^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)} X_i}{\sum_i Z_{i,1}^{(t)}}}$$
$$\overset{set}{=} 0$$

M-Step: Maximize $Q(\theta|\theta^{(t)})$ w.r.t. $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$

- $$\boxed{\pi_1^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)}}{\sum_i Z_{i,1}^{(t)} + Z_{i,2}^{(t)}} = \frac{1}{n}\sum_i Z_{i,1}^{(t)}}$$

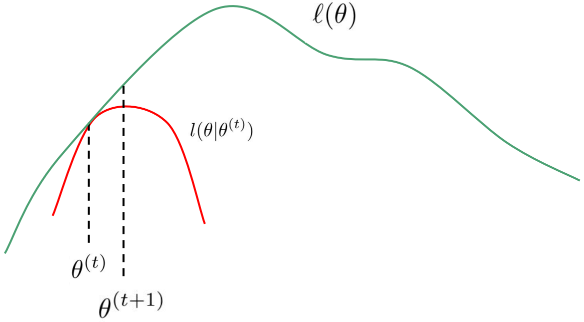
- $$\boxed{\mu_1^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)} X_i}{\sum_i Z_{i,1}^{(t)}}}$$

- σ_2: $$\frac{\partial Q}{\partial \sigma_1} = \frac{\partial}{\partial \sigma_1} \sum_{i=1}^{n} \left(-\log \sigma_1 - \frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) Z_{i,1}^{(t)}$$
$$= \sum_{i=1}^{n} \left(-\frac{1}{\sigma_1} + \frac{(x_i - \mu_1)^2}{\sigma_1^3} \right) Z_{i,1}^{(t)} \implies \boxed{(\sigma_1^2)^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)} (X_i - \mu_1^{(t+1)})^2}{\sum_i Z_{i,1}^{(t)}}}$$
$$\stackrel{set}{=} 0$$

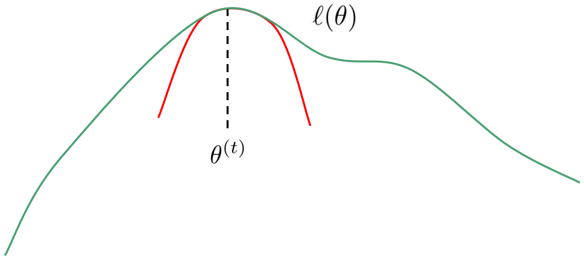


M-Step: Maximize $Q(\theta|\theta^{(t)})$ w.r.t. $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$

- $$\pi_1^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)}}{\sum_i Z_{i,1}^{(t)} + Z_{i,2}^{(t)}} = \frac{1}{n}\sum_i Z_{i,1}^{(t)}$$
- $$\mu_1^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)} X_i}{\sum_i Z_{i,1}^{(t)}}$$
- $$(\sigma_1^2)^{(t+1)} = \frac{\sum_i Z_{i,1}^{(t)}(X_i - \mu_1^{(t+1)})^2}{\sum_i Z_{i,1}^{(t)}}$$
- Similar process for μ_2 & σ_2



Repeat this process until we find a (local) maximum



Go to `week8/em_lab.R`