10-701 Introduction to Machine Learning

The EM Algorithm

Spring 2019

Ameet Talwalkar
(slide credit: Virginia Smith)

Outline

1. Gaussian mixture models

2. GMMs and Incomplete Data

3. EM Algorithm

EM Algorithm

EM algorithm: motivation and setup

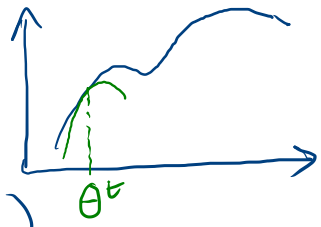
- EM is a general procedure to estimate parameters for probabilistic models with hidden/latent variables
- Suppose the model is given by a joint distribution

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$$

$p(x, z; \theta)$

$\mathcal{D} = \{x_n\}$: x : incomplete data

model: $p(x; \theta) = \sum_z p(x, z; \theta)$



$$l(\theta) = \log \prod_n p(x_n; \theta)$$

$$= \sum_n \log p(x_n; \theta)$$

$$= \sum_n \log \sum_{z_n} p(x_n, z_n; \theta)$$

$$= \sum_n \log \sum_{z_n} q_n(z_n) \frac{p(x_n, z_n; \theta)}{q_n(z_n)}$$

$$= \sum_n \log E_q\left(\frac{p(x_n, z_n; \theta)}{q_n(z_n)}\right)$$

$$\geq \sum_n E_q\left(\log\left(\frac{p(x_n, z_n; \theta)}{q_n(z_n)}\right)\right)$$

Jensen's Inequality

f: strictly concave

(i) $f(E(x)) \geq E(f(x))$

(ii) equality will hold if x is constant r.v.

$$h(\theta) = E_x\{g(x, \theta)\}$$

Equality holds if: $\frac{p(x_n, z_n; \theta^t)}{q_n(z_n)} = \text{const}$ $\forall$ values z_n

$\Rightarrow q_n(z_n) \propto p(x_n, z_n; \theta^t)$

$\Rightarrow q_n(z_n) = c\, p(x_n, z_n; \theta^t)$ $\forall$ values of z_n

$\Rightarrow \sum_{z_n} q_n(z_n) = 1 = c \sum_{z_n} p(x_n; z_n; \theta^t)$

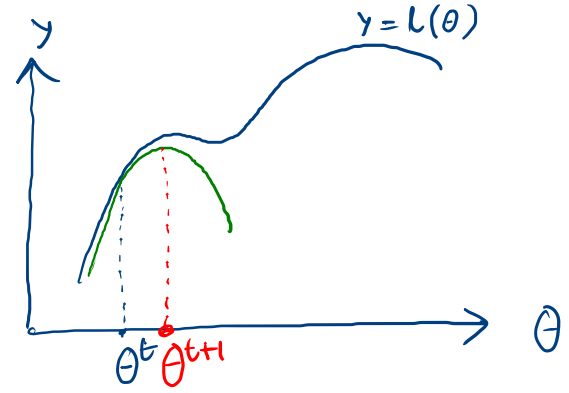
$= c\, p(x_n; \theta^t)$

$\Rightarrow c = \frac{1}{p(x_n; \theta^t)}$

$$q_n(z_n) = \frac{p(x_n, z_n; \theta^t)}{p(x_n; \theta^t)} = p(z_n | x_n; \theta^t)$$

$$l(\theta^t) = \sum_n E_q\left(\log\left(\frac{p(x_n, z_n; \theta^t)}{p(z_n | x_n; \theta^t)}\right)\right)$$

$$l(\theta) \geq \underbrace{\sum_n \sum_{z_n} p(z_n | x_n; \theta^t) \log \frac{p(x_n, z_n; \theta)}{p(z_n | x_n; \theta^t)}}_{ELBO(\theta, \theta^t)}$$



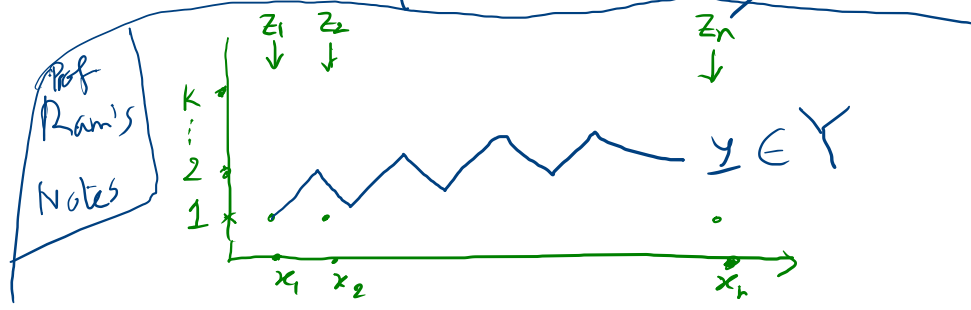
E-Step: (i) Compute the posterior $p(z_n | x_n; \theta^t)$

$$\circledast \quad Q(\theta, \theta^t) = \sum_n \sum_{z_n} p(z_n | x_n; \theta^t) \log p(x_n, z_n; \theta) = \sum_n E_q \log p(x_n, z_n; \theta)$$

$$= E_q\left(\log \prod_n (x_n, z_n; \theta)\right) \quad \leftarrow$$

M-Step:

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^t)$$



EM algorithm: motivation and setup

- EM is a general procedure to estimate parameters for probabilistic models with hidden/latent variables
- Suppose the model is given by a joint distribution

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$$

- Given **incomplete data** $\mathcal{D} = \{\boldsymbol{x}_n\}$ our goal is to compute MLE of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \arg\max \ell(\theta) = \arg\max \log \mathcal{D} = \arg\max \sum_n \log p(\boldsymbol{x}_n|\boldsymbol{\theta})$$

$$= \arg\max \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$$

The objective function $\ell(\boldsymbol{\theta})$ is called *incomplete* log-likelihood

A lower bound

- log-sum form of incomplete log-likelihood is difficult to work with
- EM: construct lower bound on $\ell(\boldsymbol{\theta})$ (E-step) and optimize it (M-step)

A lower bound

- log-sum form of incomplete log-likelihood is difficult to work with
- EM: construct lower bound on $\ell(\boldsymbol{\theta})$ (E-step) and optimize it (M-step)
- If we define $q(\boldsymbol{z})$ as a distribution over $\boldsymbol{z}$, then

$$\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\theta})$$

A lower bound

- log-sum form of incomplete log-likelihood is difficult to work with
- EM: construct lower bound on $\ell(\boldsymbol{\theta})$ (E-step) and optimize it (M-step)
- If we define $q(\boldsymbol{z})$ as a distribution over $\boldsymbol{z}$, then

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) \\ &= \sum_n \log \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}\end{aligned}$$

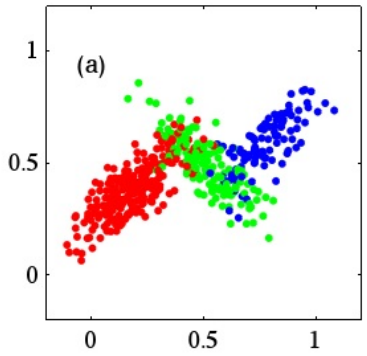
A lower bound

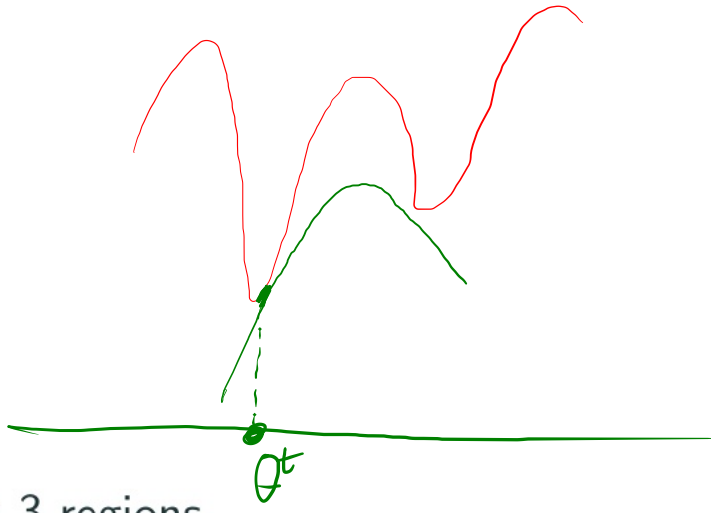
- log-sum form of incomplete log-likelihood is difficult to work with
- EM: construct lower bound on $\ell(\boldsymbol{\theta})$ (E-step) and optimize it (M-step)
- If we define $q(\boldsymbol{z})$ as a distribution over $\boldsymbol{z}$, then

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) \\ &= \sum_n \log \sum_{z_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \\ &\geq \sum_n \sum_{z_n=1}^{K} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}\end{aligned}$$

- Last step follows from Jensen's inequality, i.e., $f(\mathbb{E}X) \geq \mathbb{E}f(X)$ for concave function f

GMM Example





- Consider the previous model where $\boldsymbol{x}$ could be from 3 regions
- We can choose $q(\boldsymbol{z})$ as any valid distribution
- e.g., $q(z = k) = 1/3$ for any of 3 colors
- e.g., $q(z = k) = 1/2$ for red and blue, 0 for green

Which $q(\boldsymbol{z})$ should we choose?

Which $q(z)$ to choose?

Recall:

$$\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \sum_n \log \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

$$\geq \sum_n \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

- The lower bound we derived for $\ell(\boldsymbol{\theta})$ holds for all choices of $q(\cdot)$
- We want a *tight* lower bound

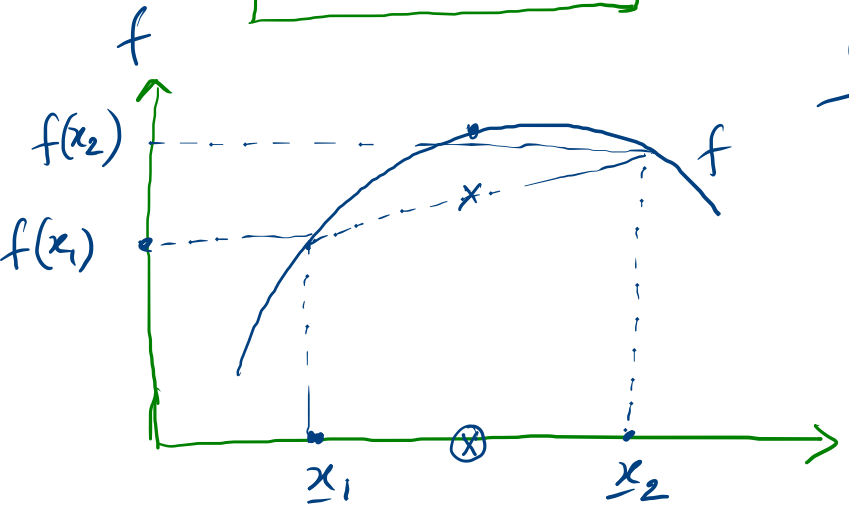
Which $q(z)$ to choose?

Recall:

$$\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

$$\geq \sum_n \sum_{z_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

- The lower bound we derived for $\ell(\boldsymbol{\theta})$ holds for all choices of $q(\cdot)$
- We want a *tight* lower bound, and given some current estimate $\boldsymbol{\theta}^t$, we will pick $q_n(\cdot)$ such that our lower bound holds *with equality* at $\boldsymbol{\theta}^t$
- $f(\mathbb{E}X) = \mathbb{E}f(X)$?



Concavity:

$$f(t\underline{x}_1 + (1-t)\underline{x}_2) \geq t\, f(\underline{x}_1) + (1-t)\, f(\underline{x}_2)$$

Which $q(z)$ to choose?

Recall:

$$\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

$$\geq \sum_n \sum_{z_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

- The lower bound we derived for $\ell(\boldsymbol{\theta})$ holds for all choices of $q(\cdot)$
- We want a *tight* lower bound, and given some current estimate $\boldsymbol{\theta}^t$, we will pick $q_n(\cdot)$ such that our lower bound holds *with equality* at $\boldsymbol{\theta}^t$
- $f(\mathbb{E}X) = \mathbb{E}f(X)$? It is sufficient for X to be a constant random variable!

Which $q(z)$ to choose?

Recall:

$$\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

$$\geq \sum_n \sum_{z_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

- The lower bound we derived for $\ell(\boldsymbol{\theta})$ holds for all choices of $q(\cdot)$
- We want a *tight* lower bound, and given some current estimate $\boldsymbol{\theta}^t$, we will pick $q_n(\cdot)$ such that our lower bound holds *with equality* at $\boldsymbol{\theta}^t$
- $f(\mathbb{E}X) = \mathbb{E}f(X)$? It is sufficient for X to be a constant random variable!
- Choose $q_n(\boldsymbol{z}_n) \propto p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)$!

Which $q(z)$ to choose?

Recall:

$$\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

$$\geq \sum_n \sum_{z_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} = c$$

$\Rightarrow q_n(z_n) = \frac{1}{c} p(x_n, z_n; \theta)$

$\Rightarrow \sum_{z_n} q_n(z_n) = 1 = \frac{1}{c} \sum_{z_n} p(x_n, z_n; \theta)$

$\Rightarrow c = \sum_{z_n} p(x_n, z_n, \theta)$

- The lower bound we derived for $\ell(\boldsymbol{\theta})$ holds for all choices of $q(\cdot)$
- We want a *tight* lower bound, and given some current estimate $\boldsymbol{\theta}^t$, we will pick $q_n(\cdot)$ such that our lower bound holds *with equality* at $\boldsymbol{\theta}^t$
- $f(\mathbb{E}X) = \mathbb{E}f(X)$? It is sufficient for X to be a constant random variable!
- Choose $q_n(\boldsymbol{z}_n) \propto p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)$! Since $q_n(\cdot)$ is a distribution, we have

$$q_n(\boldsymbol{z}_n) = \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)}{\sum_k p(\boldsymbol{x}_n, z_n = k|\boldsymbol{\theta}^t)} = \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)}{p(\boldsymbol{x}_n|\boldsymbol{\theta}^t)} = p(\boldsymbol{z}_n|\boldsymbol{x}_n; \boldsymbol{\theta}^t)$$

Which $q(z)$ to choose?

Recall:

$$\ell(\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \sum_n \log \sum_{z_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

$$\geq \sum_n \sum_{z_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)}$$

- The lower bound we derived for $\ell(\boldsymbol{\theta})$ holds for all choices of $q(\cdot)$
- We want a *tight* lower bound, and given some current estimate $\boldsymbol{\theta}^t$, we will pick $q_n(\cdot)$ such that our lower bound holds *with equality* at $\boldsymbol{\theta}^t$
- $f(\mathbb{E}X) = \mathbb{E}f(X)$? It is sufficient for X to be a constant random variable!
- Choose $q_n(\boldsymbol{z}_n) \propto p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)$! Since $q_n(\cdot)$ is a distribution, we have

$$q_n(\boldsymbol{z}_n) = \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)}{\sum_k p(\boldsymbol{x}_n, z_n = k|\boldsymbol{\theta}^t)} = \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)}{p(\boldsymbol{x}_n|\boldsymbol{\theta}^t)} = p(\boldsymbol{z}_n|\boldsymbol{x}_n; \boldsymbol{\theta}^t)$$

- This is the **posterior distribution** of z_n given $\boldsymbol{x}_n$ and $\boldsymbol{\theta}^t$

E and M Steps

Our simplified expression

$$\ell(\boldsymbol{\theta}^t) = \sum_n \sum_{z_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)}$$

E and M Steps

Our simplified expression

$$\ell(\boldsymbol{\theta}^t) = \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)}$$

E-Step: For all n, compute $q_n(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)$

Why is this called the E-Step?

E and M Steps

Our simplified expression

$$\ell(\boldsymbol{\theta}^t) = \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)}$$

E-Step: For all n, compute $q_n(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)$

Why is this called the E-Step? Because we can view it as computing the *expected (complete) log-likelihood*:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \mathbb{E}_q \sum_n \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$$

E and M Steps

Our simplified expression

$$\ell(\boldsymbol{\theta}^t) = \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}^t)}{p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)}$$

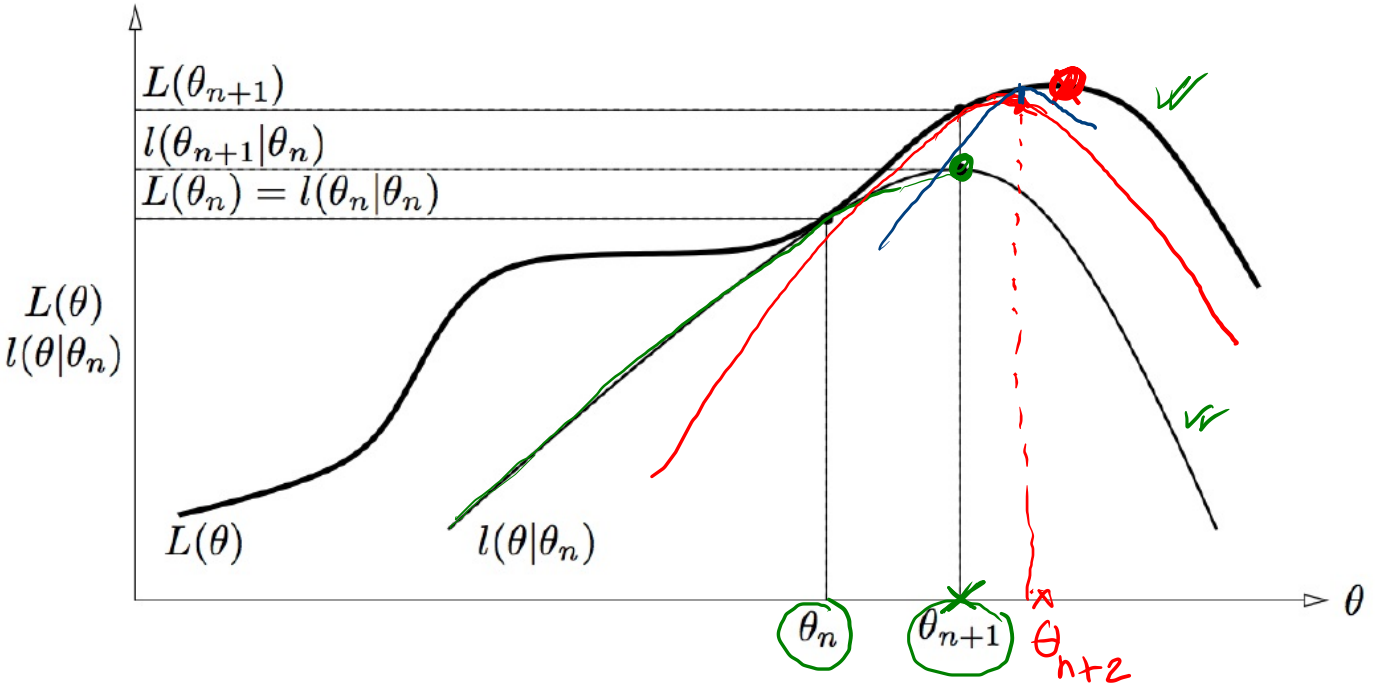
E-Step: For all n, compute $q_n(\boldsymbol{z}_n) = p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t)$

Why is this called the E-Step? Because we can view it as computing the *expected (complete) log-likelihood*:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_n \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n;\boldsymbol{\theta}^t) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}) = \mathbb{E}_q \sum_n \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta})$$

M-Step: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$, i.e., $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$

EM in Pictures



(Figure from tutorial by Sean Borman)

Iterative and monotonic improvement

- We can show that $\ell(\boldsymbol{\theta}^{t+1}) \geq \ell(\boldsymbol{\theta}^{t})$

Iterative and monotonic improvement

- We can show that $\ell(\boldsymbol{\theta}^{t+1}) \geq \ell(\boldsymbol{\theta}^{t})$
- Recall that we chose $q(\cdot)$ in the E-step such that:

$$\ell(\boldsymbol{\theta}^{t}) = \sum_{n} \sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\theta}^{t})}{q(\boldsymbol{z}_n)}$$

Iterative and monotonic improvement

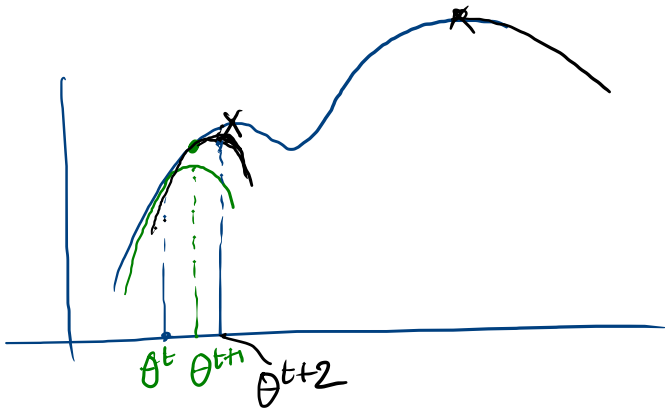
- We can show that $\ell(\boldsymbol{\theta}^{t+1}) \geq \ell(\boldsymbol{\theta}^t)$
- Recall that we chose $q(\cdot)$ in the E-step such that:

$$\ell(\boldsymbol{\theta}^t) = \sum_n \sum_{z_n} q(z_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\theta}^t)}{q(\boldsymbol{z}_n)}$$

- However, in the M-step, $\boldsymbol{\theta}^{t+1}$ is chosen to maximize the right hand side of the equation, thus proving our desired result

$$\ell(\theta^{t+1}) \geq \text{ELBO}(\theta, \theta^{(t+1)})$$
$$\geq \ell(\theta^t)$$

Iterative and monotonic improvement



- We can show that $\ell(\boldsymbol{\theta}^{t+1}) \geq \ell(\boldsymbol{\theta}^t)$
- Recall that we chose $q(\cdot)$ in the E-step such that:

$$\ell(\boldsymbol{\theta}^t) = \sum_n \sum_{z_n} q(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n | \boldsymbol{\theta}^t)}{q(\boldsymbol{z}_n)}$$

- However, in the M-step, $\boldsymbol{\theta}^{t+1}$ is chosen to maximize the right hand side of the equation, thus proving our desired result
- Note: the EM procedure converges but only to a local optimum

You should know . . .

- EM is a general procedure for maximizing a likelihood with *latent (unobserved) variables*

You should know . . .

- EM is a general procedure for maximizing a likelihood with *latent (unobserved) variables*
- The two steps of EM:

You should know . . .

- EM is a general procedure for maximizing a likelihood with *latent (unobserved) variables*
- The two steps of EM:
  - (1) Estimating unobserved data from observed data and current parameters

You should know . . .

- EM is a general procedure for maximizing a likelihood with *latent (unobserved) variables*
- The two steps of EM:
  - (1) Estimating unobserved data from observed data and current parameters
  - (2) Using this "complete" data to find the maximum likelihood parameter estimates

You should know . . .

- EM is a general procedure for maximizing a likelihood with *latent (unobserved) variables*
- The two steps of EM:
  - (1) Estimating unobserved data from observed data and current parameters
  - (2) Using this "complete" data to find the maximum likelihood parameter estimates
- Pros: Guaranteed to converge, no parameters to tune (e.g., compared to gradient methods)

You should know . . .

- EM is a general procedure for maximizing a likelihood with *latent (unobserved) variables*
- The two steps of EM:
  - (1) Estimating unobserved data from observed data and current parameters
  - (2) Using this "complete" data to find the maximum likelihood parameter estimates
- Pros: Guaranteed to converge, no parameters to tune (e.g., compared to gradient methods)
- Cons: Can get stuck in local optima, can be expensive

You should know . . .

- EM is a general procedure for maximizing a likelihood with *latent (unobserved) variables*
- The two steps of EM:
  - (1) Estimating unobserved data from observed data and current parameters
  - (2) Using this "complete" data to find the maximum likelihood parameter estimates
- Pros: Guaranteed to converge, no parameters to tune (e.g., compared to gradient methods)
- Cons: Can get stuck in local optima, can be expensive
- Why is EM useful for unsupervised learning?

You should know . . .

- EM is a general procedure for maximizing a likelihood with *latent (unobserved) variables*
- The two steps of EM:
  - (1) Estimating unobserved data from observed data and current parameters
  - (2) Using this "complete" data to find the maximum likelihood parameter estimates
- Pros: Guaranteed to converge, no parameters to tune (e.g., compared to gradient methods)
- Cons: Can get stuck in local optima, can be expensive
- Why is EM useful for unsupervised learning?
  - EM is a general method to deal with hidden data; we have studied it in the context of hidden *labels* (unsupervised learning)