

Lecture 23. Expectation Maximization.

COMP90051 Statistical Machine Learning

Semester 2, 2020
Lecturer: Ben Rubinstein



THE UNIVERSITY OF
MELBOURNE

This lecture

- EM intuition by GMM with recap
- Lower-bound $\log p(\mathbf{X}|\boldsymbol{\theta})$ by $\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z})]$
 - * Holds for any $\boldsymbol{\theta}, p(\mathbf{Z})$
 - * Uses Jensen's inequality (concavity of log)
- Maximise not $\log p(\mathbf{X}|\boldsymbol{\theta})$ but lower bound, alternating:
 - * E-Step: choose $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*)$ raises lower bound up to log-likelihood, for any $\boldsymbol{\theta}^*$
 - * M-Step: $\boldsymbol{\theta}^*$ by max'ing "completed" log-likelihood; ideally, easy MLE
- Proving the E-step
- Applying EM to GMM MLE-based training
 - * Coming full circle back to comparison to k -means

Recap: EM for fitting the GMM

- Initialisation Step:

- * Initialize K clusters: C_1, \dots, C_K
 (μ_j, Σ_j) and $P(C_j)$ for each cluster j .

- Iteration Step:

- * Estimate the cluster of each datum

$$p(C_j | x_i)$$

Calculating the probability for each point that for each cluster



Expectation

- * Re-estimate the cluster parameters



Maximisation

$$(\mu_j, \Sigma_j), p(C_j) \text{ for each cluster } j$$

Re-calculating the parameter for each cluster to maximise the likelihood of the points

MLE vs EM

- MLE is a frequentist *principle* that suggests that given a dataset, the “best” parameters to use are the ones that maximise the probability of the data
 - * MLE is a way *to formally pose* the problem
- EM is an *algorithm*
 - * EM is a way *to solve* the problem posed by MLE
 - * Especially convenient under unobserved latent variables
- MLE can be found by other methods such as gradient descent (but gradient descent is not always the most convenient method)

EM for GMM and generally

- EM is a general approach, goes beyond GMMs
 - * Purpose: Implement **MLE under latent variables \mathbf{Z}** ('latent' is fancy for 'missing')
- What are variables, parameters in GMMs?
 - * Variables: Point locations \mathbf{X} and cluster assignments \mathbf{Z}
 - let z_i denote true cluster membership for each point x_i , computing the likelihood with known values \mathbf{z} is simplified (see next section)
 - * Parameters: $\boldsymbol{\theta}$ are cluster locations and scales
- What is EM really doing?
 - * **Coordinate ascent** on a lower bound on the log-likelihood
 - M-step: ascent in modeled parameters $\boldsymbol{\theta}$ Ascending on the cluster locations and scales
 - E-step: ascent in the marginal likelihood $P(\mathbf{Z})$ Ascending on the membership for each point
 - * Each step moves towards a **local** optimum
 - * Can get stuck, can need **random restarts**

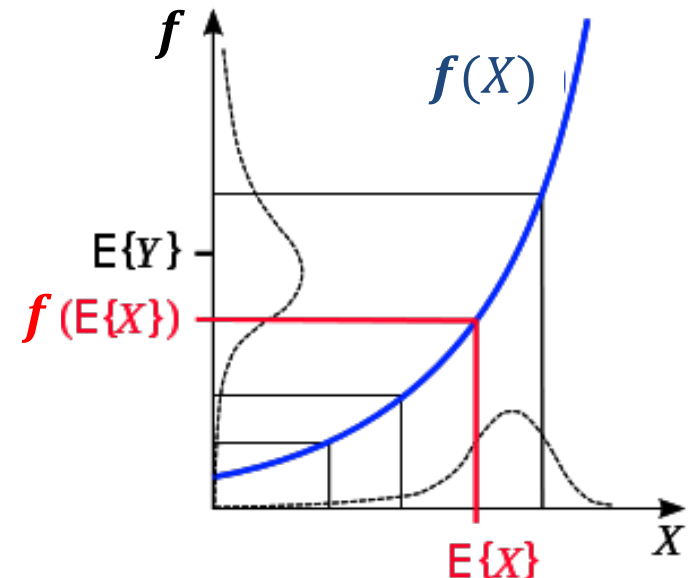
Using convexity: Jensen's inequality

- Compares effect of averaging before and after applying a **convex function**:

$$f(\text{Average}(\mathbf{x})) \leq \text{Average}(f(\mathbf{x}))$$
- Example:
 - * Let f be some convex function, such as $f(x) = x^2$
 - * Consider $\mathbf{x} = [1, 2, 3, 4, 5]'$, then $f(\mathbf{x}) = [1, 4, 9, 16, 25]'$
 - * Average of input $\text{Average}(\mathbf{x}) = 3$
 - * $f(\text{Average}(\mathbf{x})) = 9$
 - * Average of output $\text{Average}(f(\mathbf{x})) = 12.4$
- Proof follows from the definition of convexity
 - * Proof by induction

General statement:

- * If \mathbf{X} random variable, f is a convex function
- * $f(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[f(\mathbf{X})]$



Putting the latent variables to use

We want to maximise $\log p(\mathbf{X}|\boldsymbol{\theta})$. We don't observe \mathbf{Z} (here discrete), but can (re)introduce it nonetheless.

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

$$= \log \sum_{\mathbf{Z}} \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \frac{p(\mathbf{Z})}{p(\mathbf{Z})} \right)$$

$$= \log \sum_{\mathbf{Z}} \left(\overset{p_r}{p(\mathbf{Z})} \overset{\times}{\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z})}} \right) \xrightarrow{\text{definition of expectation}}$$

$$= \log \mathbb{E}_{\mathbf{Z}} \left[\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z})} \right]$$

$$\geq \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z})} \right]$$

$$= \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{Z})]$$

← Marginalisation (here $\sum_{\mathbf{Z}} \dots$ iterates over all possible values of \mathbf{Z})

← Need \mathbf{Z} to have non-zero marginal

Example of an marginal probability

$\mathbf{Y} \backslash \mathbf{X}$	x_1	x_2	x_3	x_4	$p_Y(\mathbf{y}) \downarrow$
y_1	4/32	2/32	1/32	1/32	8/32
y_2	3/32	6/32	3/32	3/32	15/32
y_3	9/32	0	0	0	9/32
$p_X(\mathbf{x}) \rightarrow$	16/32	8/32	4/32	4/32	32/32

← Jensen's inequality holds in this direction since $\log(\dots)$ is a concave function

Since we move a convex function to the inside

Maximising the lower bound (1/2)

- $\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z})]$
- The right hand side (RHS) is a **lower bound** on the original log likelihood
 - * This holds for any $\boldsymbol{\theta}$ and any non zero $p(\mathbf{Z})$
- Intuitively, we want to push the lower bound up
- This lower bound is a function of **two “variables” $\boldsymbol{\theta}$ and $p(\mathbf{Z})$** . We want to maximise the RHS as a function of these two “variables”
- It is hard to optimise with respect to both at the same time, so EM resorts to an iterative procedure

Maximising the lower bound (2/2)

- $\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z})]$

- EM is essentially **coordinate ascent**:

- * Fix $\boldsymbol{\theta}$ and optimise the lower bound for $p(\mathbf{Z})$
- * Fix $p(\mathbf{Z})$ and optimise for $\boldsymbol{\theta}$

we will
prove this
shortly

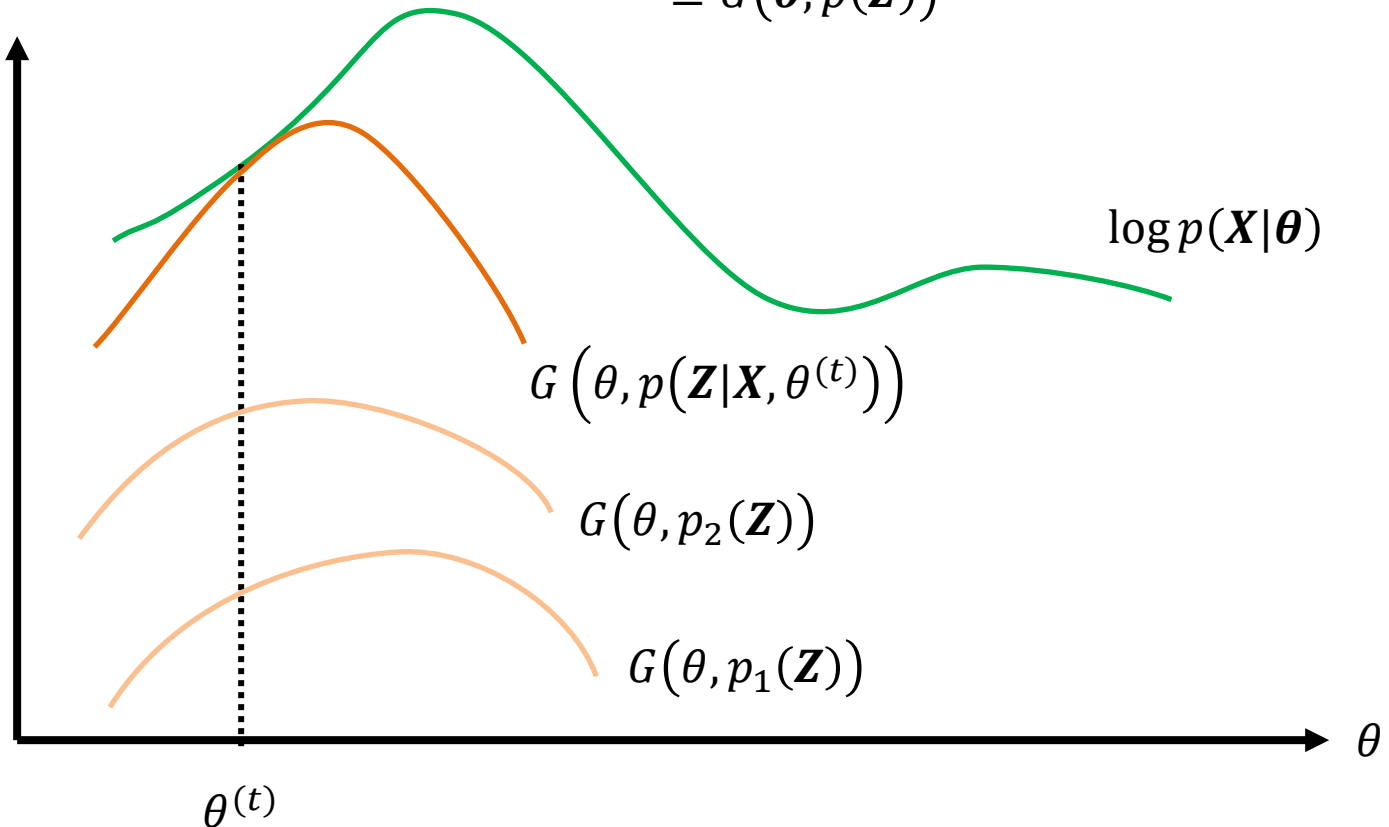
- The convenience of EM comes from the following

- For any point $\boldsymbol{\theta}^*$, it can be shown that setting $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*)$ makes the lower bound tight

- For any $p(\mathbf{Z})$, the second term does not depend on $\boldsymbol{\theta}$
- When $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*)$, the first term can usually be maximised as a function of $\boldsymbol{\theta}$ in a closed-form
 - * If not, then probably don't use EM

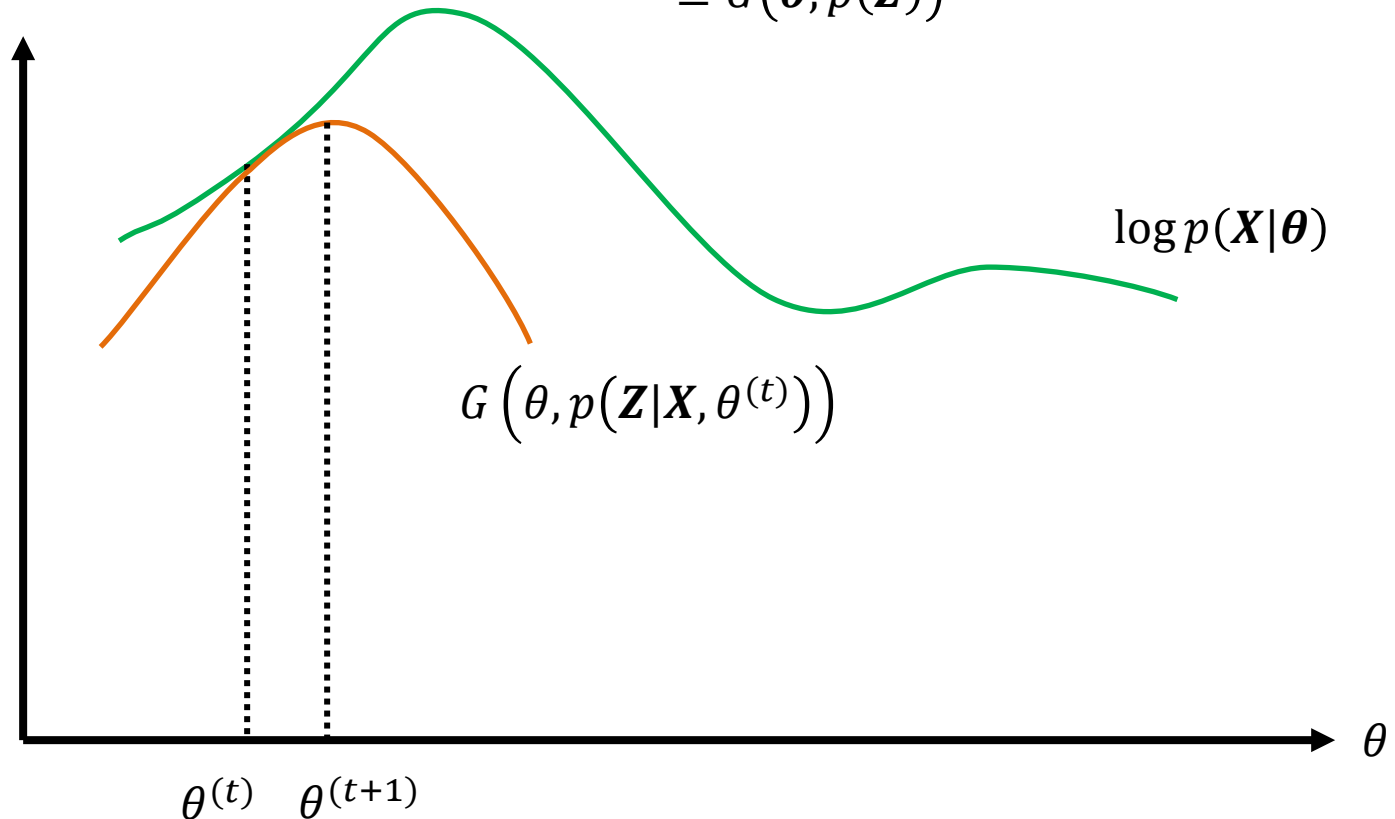
Example (1/3)

$$\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z})]}_{\equiv G(\boldsymbol{\theta}, p(\mathbf{Z}))}$$



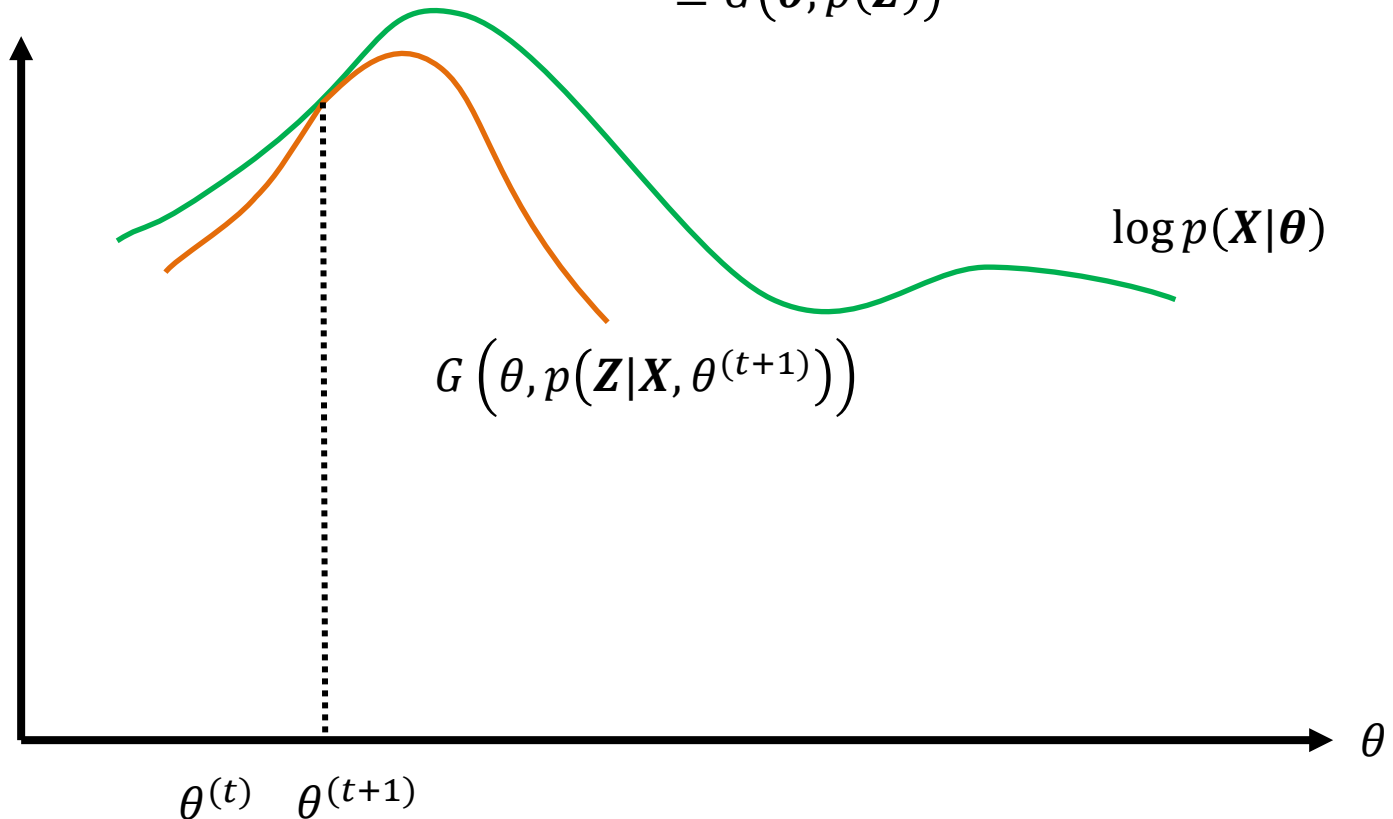
Example (2/3)

$$\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z})]}_{\equiv G(\boldsymbol{\theta}, p(\mathbf{Z}))}$$



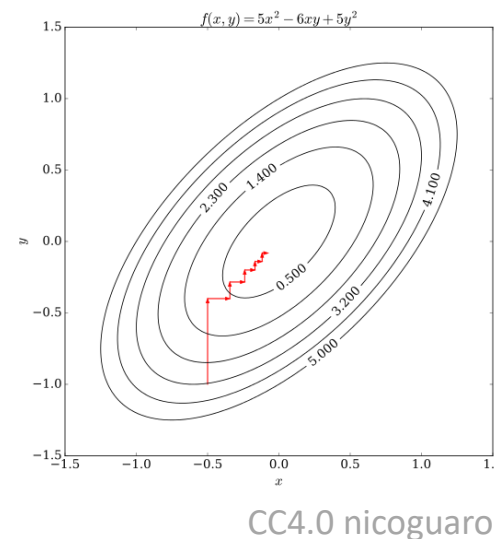
Example (3/3)

$$\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z})]}_{\equiv G(\boldsymbol{\theta}, p(\mathbf{Z}))}$$



Mini Summary

- EM intuition by GMM with recap
- Lower-bound $\log p(\mathbf{X}|\boldsymbol{\theta})$ by $\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z})]$
 - * Holds for any $\boldsymbol{\theta}, p(\mathbf{Z})$
 - * Uses Jensen's inequality (concavity of log)
- Maximise not $\log p(\mathbf{X}|\boldsymbol{\theta})$ but lower bound, alternating:
 - * E-Step: choose $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*)$ raises lower bound up to log-likelihood, for any $\boldsymbol{\theta}^*$ assign probability to each point
 - * M-Step: $\boldsymbol{\theta}^*$ by max'ing “completed” log-likelihood; ideally, easy MLE Re-locate the cluster
- The E- and M-steps implement **coordinate ascent**



Next: Proving the E-step

EM as iterative (coordinate) ascent

1. Initialisation: choose (random) initial values of $\theta^{(1)}$

2. Update:

* **E-step**: compute $Q(\theta, \theta^{(t)}) \equiv \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta^{(t)}} [\log p(\mathbf{X}, \mathbf{Z}|\theta)]$


* **M-step**: $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta, \theta^{(t)})$

3. Termination: if no change then stop

4. Go to Step 2

This algorithm will eventually stop (converge), but the resulting estimate can be only a local maximum

Maximising the lower bound (2/2)

- $\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{Z})]$
 - EM is essentially coordinate descent:
 - * Fix $\boldsymbol{\theta}$ and optimise the lower bound for $p(\mathbf{Z})$
 - * Fix $p(\mathbf{Z})$ and optimise for $\boldsymbol{\theta}$
 - The convenience of EM follows from the following
 - For any point $\boldsymbol{\theta}^*$, it can be shown that setting $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*)$ makes the lower bound tight
 - For any $p(\mathbf{Z})$, the second term does not depend on $\boldsymbol{\theta}$
 - When $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*)$, the first term can usually be maximised as a function of $\boldsymbol{\theta}$ in a closed-form
 - * If not, then probably don't use EM
- we will prove this now
- 

Putting the latent variables in use

We want to maximise $\log p(\mathbf{X}|\boldsymbol{\theta})$. We don't know \mathbf{Z} , but consider an arbitrary non-zero distribution $p(\mathbf{Z})$

$$\boxed{\log p(\mathbf{X}|\boldsymbol{\theta})} = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

← Rule of marginal distribution
(here $\sum_{\mathbf{Z}} \dots$ iterates over all possible values of \mathbf{Z})

$$= \log \sum_{\mathbf{Z}} \left(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \frac{p(\mathbf{Z})}{p(\mathbf{Z})} \right)$$

$$= \log \sum_{\mathbf{Z}} \left(p(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z})} \right)$$

$$= \log \mathbb{E}_{\mathbf{Z}} \left[\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z})} \right]$$

← Jensen's inequality holds since $\log(\dots)$ is a concave function

$$\boxed{\geq \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z})} \right]}$$

$$= \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{Z})]$$

Setting a tight lower bound (1/2)

- $$\begin{aligned}\log p(\mathbf{X}|\boldsymbol{\theta}) &\geq \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z})} \right] \\&= \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})}{p(\mathbf{Z})} \right] && \leftarrow \text{Chain rule of probability} \\&= \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{Z})} + \log p(\mathbf{X}|\boldsymbol{\theta}) \right] \\&= \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{Z})} \right] + \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{X}|\boldsymbol{\theta})] && \leftarrow \text{Linearity of } \mathbb{E}[\cdot] \\&= \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{Z})} \right] + \log p(\mathbf{X}|\boldsymbol{\theta}) && \leftarrow \mathbb{E}[\cdot] \text{ of a constant} \\&&& \text{since this part does not depend on } \mathbf{Z} \text{ at all.} \\&\geq \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{Z})} \right] + \log p(\mathbf{X}|\boldsymbol{\theta})\end{aligned}$$

Setting a tight lower bound (2/2)

Ultimate aim:
maximise this

Lower bound of what
we want to maximise

$$\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \underbrace{\mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{Z})} \right]}_{\text{Lower bound}} + \log p(\mathbf{X}|\boldsymbol{\theta})$$

First, note that this term* ≤ 0

Second, note that if $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$, then

$$\mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} \left[\log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} \right] = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} [\log 1] = 0$$

For any $\boldsymbol{\theta}^*$, setting $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*)$ maximises the lower bound on $\log p(\mathbf{X}|\boldsymbol{\theta}^*)$ and makes it tight

*Negative Kullback-Leibler divergence between $p(\mathbf{Z})$ and $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$

Mini Summary

- We're wanting to maximise the lower bound
$$\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z})} \right]$$
- We've shown RHS is $\mathbb{E}_{\mathbf{Z}} \left[\log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{Z})} \right] + \log p(\mathbf{X}|\boldsymbol{\theta})$
- And that setting $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$
 - * Makes this RHS as big as possible
 - * Makes this RHS equal to $\log p(\mathbf{X}|\boldsymbol{\theta})$
 - * \rightarrow maximises the lower bound as desired!

Next: Application of EM to GMM learning

Estimating Parameters of Gaussian Mixture Model

*Classical application of the
Expectation-Maximisation algorithm.
Completes previous lecture.*

Latent variables of GMM

- Let z_1, \dots, z_n denote **true origins** of the corresponding points $\mathbf{x}_1, \dots, \mathbf{x}_n$. Each z_i is a discrete variable that takes values in $1, \dots, k$, where k is a number of clusters

- Now compare the original log likelihood

$$\log p(\mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \log \left(\sum_{c=1}^k w_c \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \right)$$

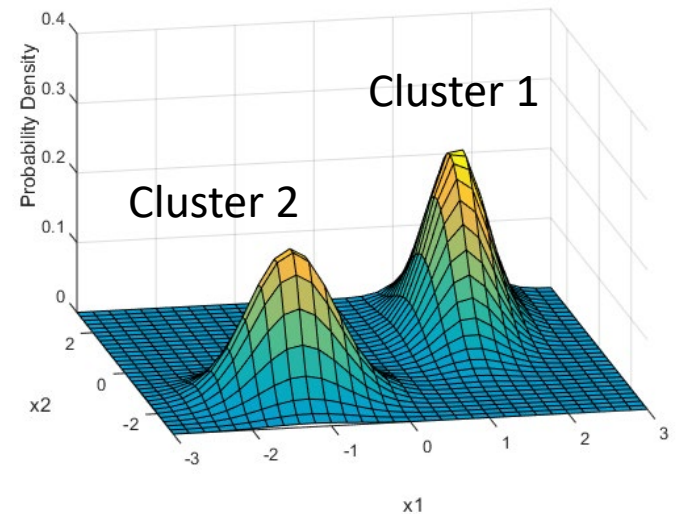
- With **complete log likelihood** (if we knew \mathbf{z})

$$\log p(\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{z}) = \sum_{i=1}^n \log \left(w_{z_i} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \right)$$

- Recall that taking a log of a normal density function results in a **tractable** expression

Handling uncertainty about \mathbf{Z}

- We cannot compute complete log likelihood because we don't know \mathbf{Z}
- EM algorithm handles this uncertainty replacing $\log p(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta})$ with expectation $\mathbb{E}_{\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})]$
- This in turn requires the distribution of $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)})$ given current parameter estimates
- Assuming that Z_i are pairwise independent, we need $P(Z_i = c|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})$
- E.g., suppose $\mathbf{x}_i = (-2, -2)$. What is the probability that this point originated from Cluster 1



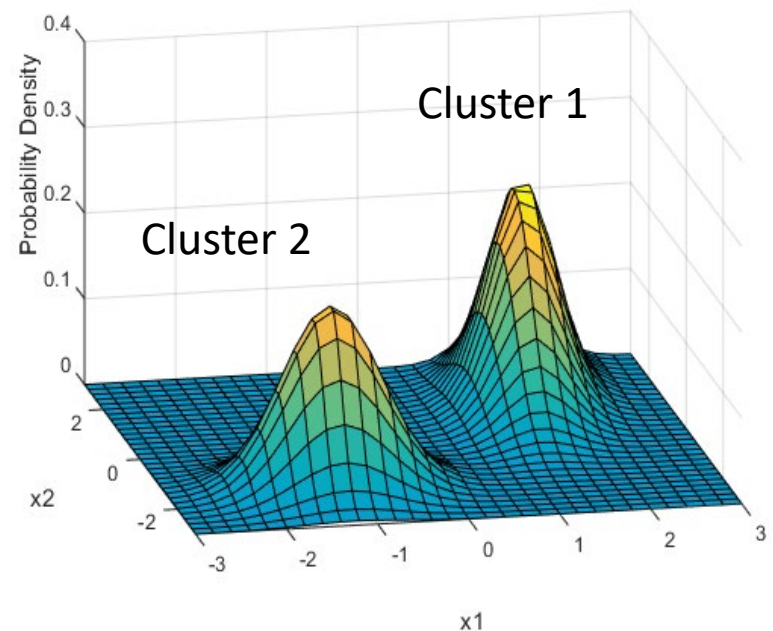
E-step: Cluster responsibilities

- Setting latent Z as originating cluster, yields (via Bayes rule)

$$P(z_i = c | \mathbf{x}_i, \boldsymbol{\theta}^{(t)}) = \frac{w_c \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{\sum_{l=1}^k w_l \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

- This probability is called *responsibility* that cluster c takes for data point i

$$r_{ic} \equiv P(z_i = c | \mathbf{x}_i, \boldsymbol{\theta}^{(t)})$$



Expectation step for GMM

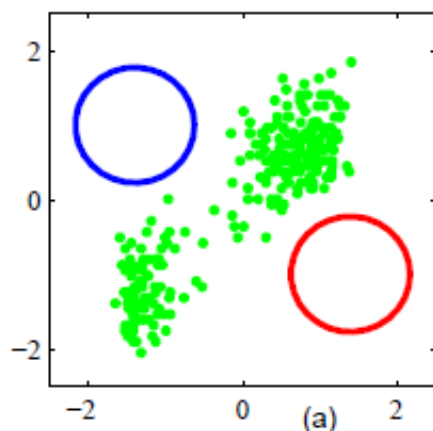
To simplify notation, we denote $\mathbf{x}_1, \dots, \mathbf{x}_n$ as \mathbf{X}

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) &\equiv \mathbb{E}_{\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}} [\log p(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta})] \\ &= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta}) \\ &= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \sum_{i=1}^n \log w_{z_i} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \\ &= \sum_{i=1}^n \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log w_{z_i} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \\ &= \sum_{i=1}^n \sum_{c=1}^k r_{ic} \log w_{z_i} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \\ &= \sum_{i=1}^n \sum_{c=1}^k r_{ic} \log w_{z_i} \\ &\quad + \sum_{i=1}^n \sum_{c=1}^k r_{ic} \log \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \end{aligned}$$

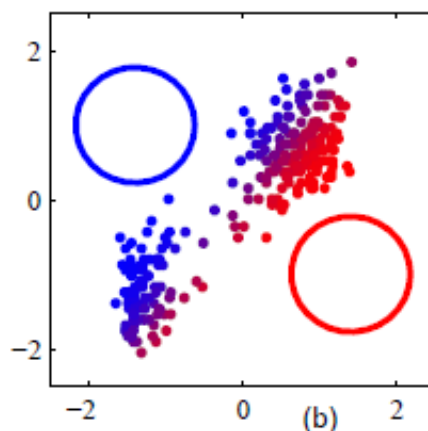
Maximisation step for GMM

- In the maximisation step, take partial derivatives of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ with respect to each of the parameters and set the derivatives to zero to obtain new parameter estimates
- $w_c^{(t+1)} = \frac{1}{n} \sum_{i=1}^n r_{ic}$
- $\boldsymbol{\mu}_c^{(t+1)} = \frac{\sum_{i=1}^n r_{ic} \mathbf{x}_i}{r_c}$
 - * Here $r_c \equiv \sum_{i=1}^n r_{ic}$
- $\boldsymbol{\Sigma}_c^{(t+1)} = \frac{\sum_{i=1}^n r_{ic} \mathbf{x}_i \mathbf{x}_i'}{r_c} - \boldsymbol{\mu}_c^{(t)} \left(\boldsymbol{\mu}_c^{(t)} \right)'$
- Note that these are the estimates for step $(t + 1)$

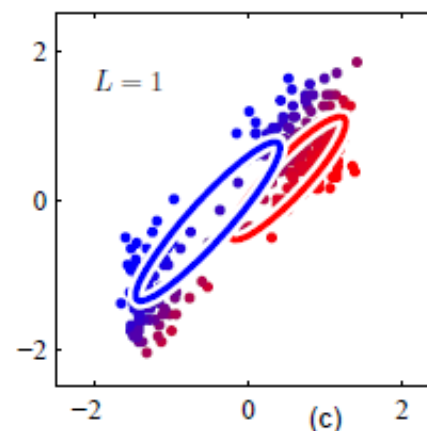
Example of fitting Gaussian Mixture model



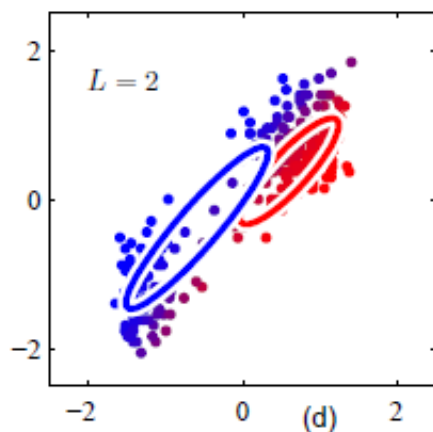
(a) Initial



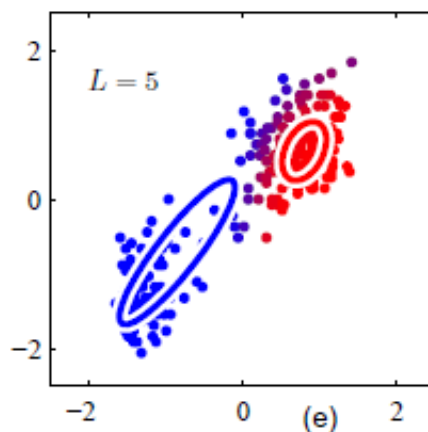
(b) E-step



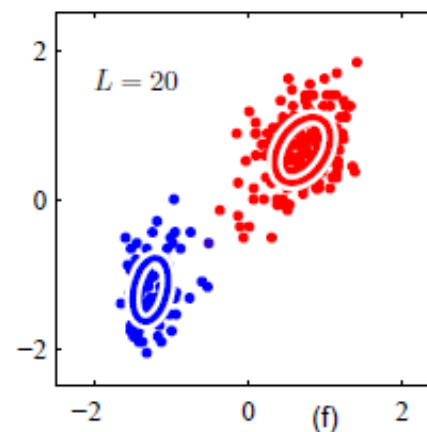
(c) M-step



(d) 2 cycles



(e) 5-cycles



(f) 20-cycles

k -means as a EM for a restricted GMM

- Consider a GMM model in which all components have the same fixed probability $w_c = 1/k$, and each Gaussian has the same fixed covariance matrix $\Sigma_c = \sigma^2 \mathbf{I}$, where \mathbf{I} is the identity matrix
- In such a model, only component centroids μ_c need to be estimated
- Next approximate a probabilistic cluster responsibility $r_{ic} = P(z_i = c | \mathbf{x}_i, \mu_c^{(t)})$ with a deterministic assignment $r_{ic} = 1$ if centroid $\mu_c^{(t)}$ is closest to point \mathbf{x}_i , and $r_{ic} = 0$ otherwise (E-step)
- Such a formulation results in a M-step where μ_c should be set as a centroid of points assigned to cluster c
- In other words, **k -means algorithm is a EM algorithm for the restricted GMM model** described above!!!

Summary

- EM as MLE algorithm under latent variables
- Maximise not $\log p(\mathbf{X}|\boldsymbol{\theta})$ but lower bound, alternating:
 - * E-Step: choose $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^*)$ raises lower bound up to log-likelihood, for any $\boldsymbol{\theta}^*$
 - * M-Step: $\boldsymbol{\theta}^*$ by max'ing “completed” log-likelihood; ideally, easy MLE
- Applying EM to GMM MLE-based training
 - * “New” interpretation of k -means as MLE with constraints