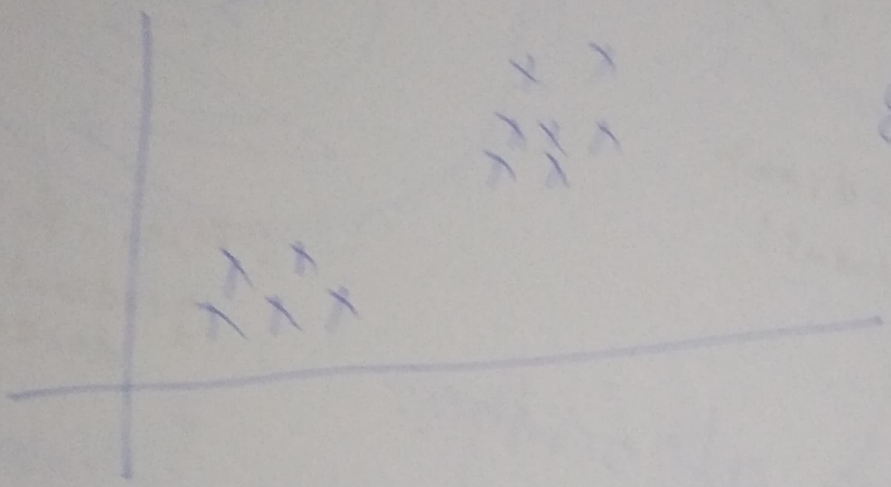
Lecture - 189

Will provide you notes later

② Unsupervised Learning:-



$\{x^{(1)}, x^{(2)} \ldots x^{(m)}\}$

K-Means Clustering:-

Data $\{x^{(1)}, \ldots x^{(m)}\}$.

1. Initialize cluster centroids.

$\mu_1 \ldots \mu_K \in \mathbb{R}^n$ randomly.

2. Repeat until convergence:

(a) Set $c^{(i)} = \arg\min_j \|x^{(i)} - \mu_j\|$

[b]: For $j = 1, \ldots k$

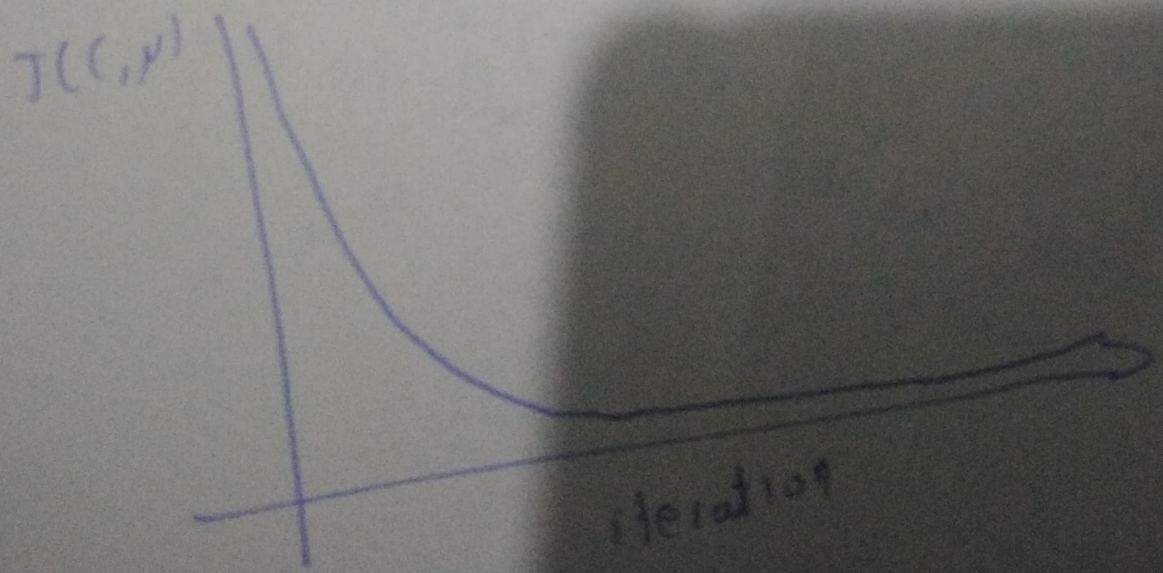
$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

(Move the cluster centroids").

$$J(c, \mu) = \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2$$

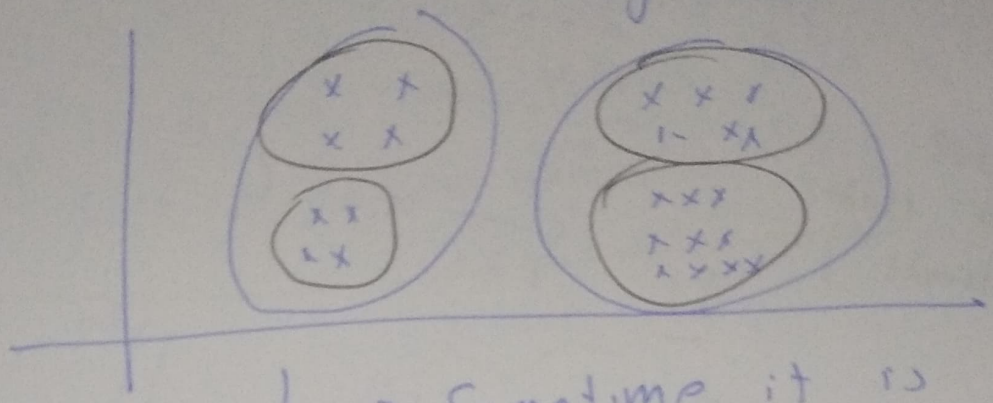
↑ "Assignments" ↑ Centroids.

On every iteration K-Means would drive down this cost. So it must converge because this function J can go below zero (it is a strictly non-negative function which goes down after every iteration).



⊛ How do you choose K?

(1) Choose K by hand C

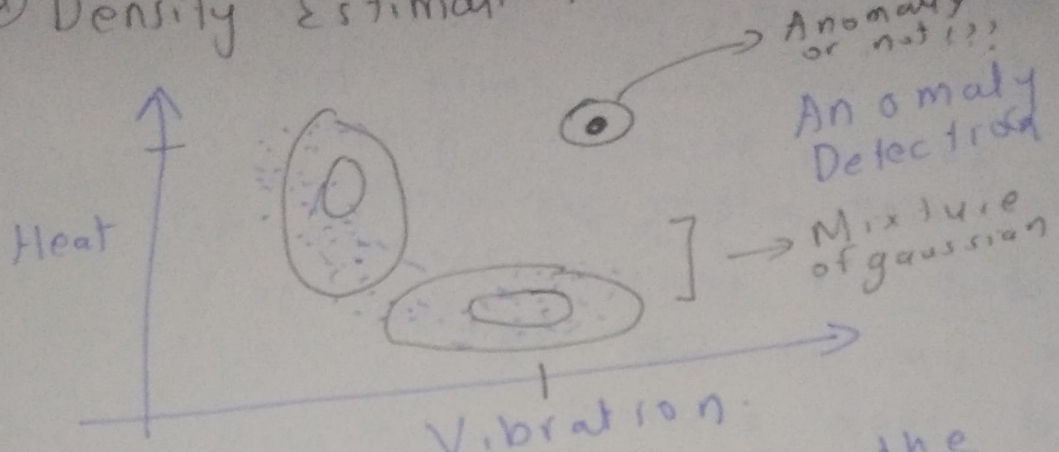


↳ Sometime it is inherently ambigious what is the right number of clusters.

(2) Criteria like AIC or BIC can be used but always try to choose k by hand depending on context, application & purpose.

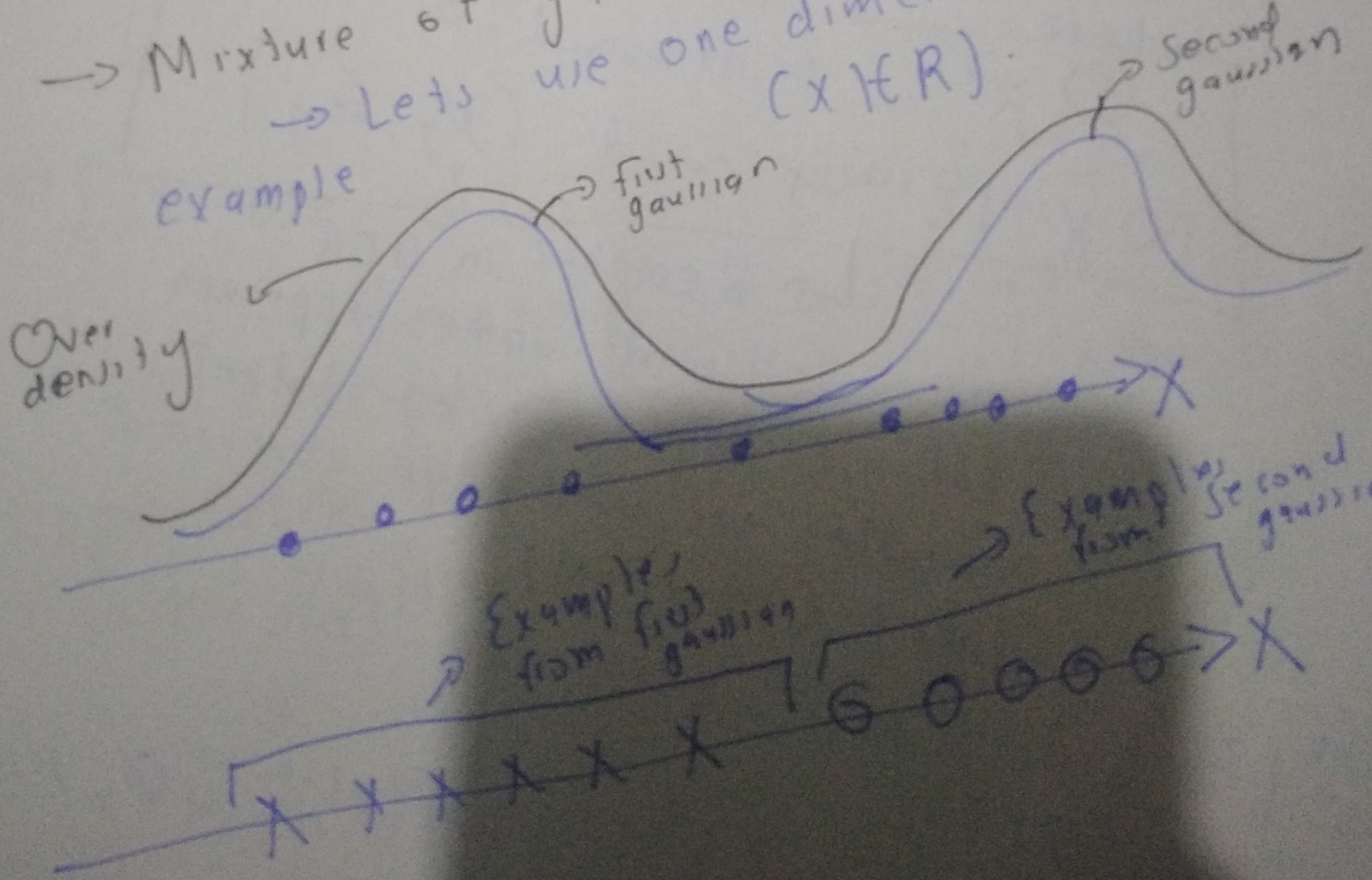
⇒ K-Mean can stuck in local minima so run it many times from different intialization and pick whichever results in the lowest cost.

⊛ Density Estimation:



↳ Model $p(x)$ :: (Modelling the density from which x was drawn
so if $p(x)$ is really small than you draw it as an anomaly

→ Mixture of gaussian model:
→ Lets use one dimension ($x \in R$).
example



In GDA model we knew labels therefore we could easily model the gaussian however, given an unsupervised setting we can't be sure about the gaussian therefore will make use of EM algorithm to model these gaussian

a) Mixture of Gaussian model:-

→ Suppose there is a latent (hidden) random variable z and $x^{(i)}, z^{(i)}$ are distributed $P(x^{(i)}, z^{(i)}) = P(x^{(i)} | z^{(i)}) \cdot P(z^{(i)})$

where

① $z^{(i)} \sim$ Multinomial (ϕ)

② $x^{(i)} | z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$

↳ Remeber in GDA we followed same approach but instead of using latent variable we used the observed label. (Important)

So → If we knew $z^{(i)}$, we can use MLE:-

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log P(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)$$

⊙ We would take the d
$\nabla_{\phi,\mu,\Sigma} \ell(\phi, \mu, \Sigma) = 0$ and would find:-

$$\phi_j = \frac{\sum_{i=1}^{m} \mathbb{1}\{z^{(i)} = j\}}{m}$$

$$\mu_j = \frac{\sum_{i=1}^{m} \mathbb{1}\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{z^{(i)} = j\}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} (x - \mu_j)^T (x - \mu_j)}{\sum_{i=1}^{m} \mathbb{1}\{z^{(i)} = j\}}$$

← Formulas for GDA as because we know z (the label).

⇝ But we don't know $z^{(i)}$'s.

→ Therefore we will use EM (Expectation minimization algorithm).

b) → Expectation Minimization Algorithm:-
(Called Bootstrap procedure):-

→ EM algorithm has two main steps:-

1) E-step (Guess value of $z^{(i)}$'s) :-

→ Set $w_j^{(i)} = P(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$

$$= \frac{\overset{①}{P(x^{(i)} | z^{(i)} = j)} \overset{②}{P(z^{(i)} = j)}}{\sum_{l=1}^{k} P(x^{(i)} | z^{(i)} = l) \cdot P(z^{(i)} = l)}$$

① $\sim \mathcal{N}(\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right)$

→ ~~$\mathcal{N}(\mu_j, \Sigma_j) =$~~

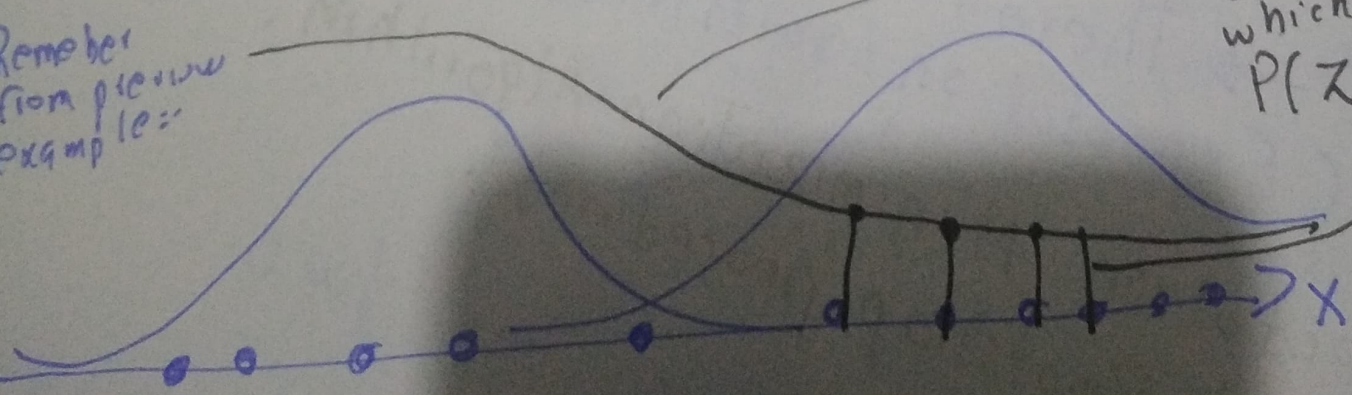
② $\sim$ Multinomial distribution $= \phi_j$

($z \sim$ Multinomially Distributed).

→ The terms in denominator would follow the same same distribution for different l's.

→ Sigmoid (Also plotted previously) curve which shows $P(z^{(i)} = j | x^{(i)})$

Remember from previous example :-



→ You could compute $w_j^{(i)}$ for every training example with every j.

2) M-Step:

$$\phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

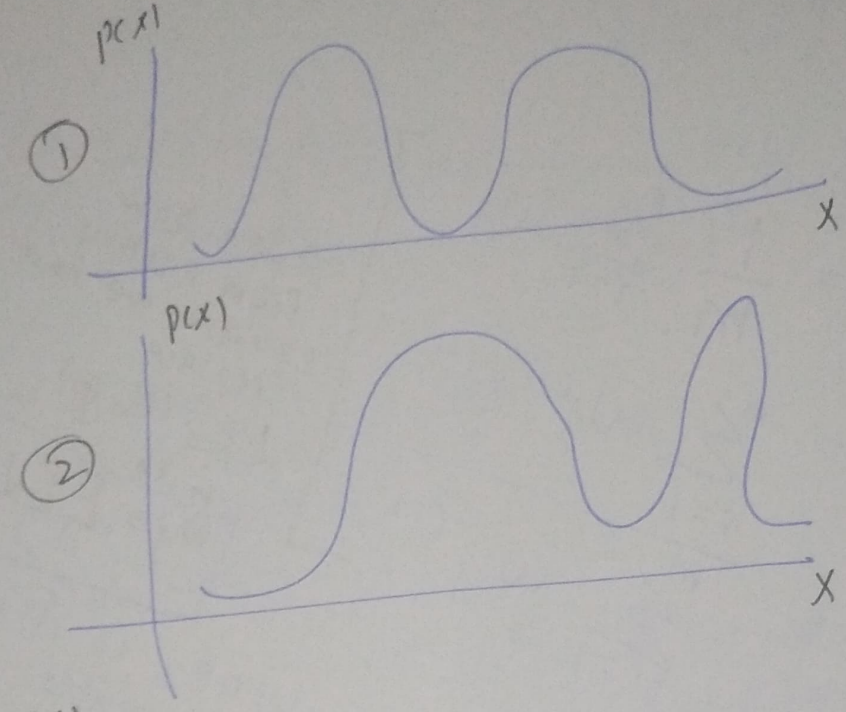
→ See how from GDA equation we have replace $1\{z^{(i)} = j\}$ with $w_j = E[1\{z^{(i)} = j\}]$

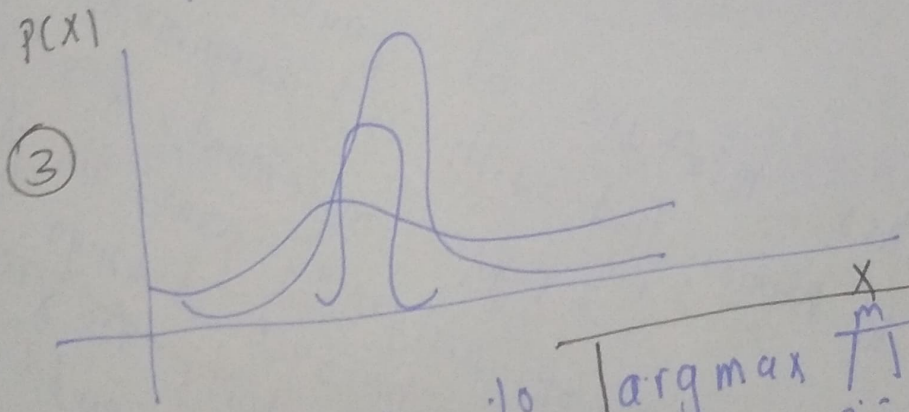
↳ this mixture of gaussian is like K-Means but with soft-assignments.

↳ EM implements a softer way of assigning points to different centroids because it assign a specific probability for $x^{(i)}$ for each j (possible centers).

→ the algorithm Mixture of Gaussian will converge but with some caveats.

→ Mixture of gaussian can fit different and various kind of models:-

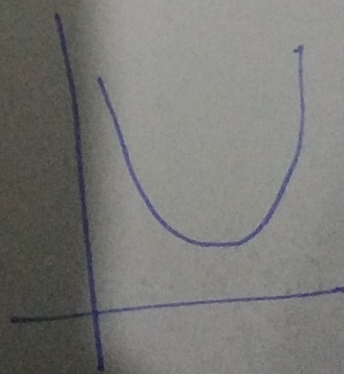




→ EM is trying to $\underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} P(x^{(i)}; \theta)$

→ Jensen's Inequality:-

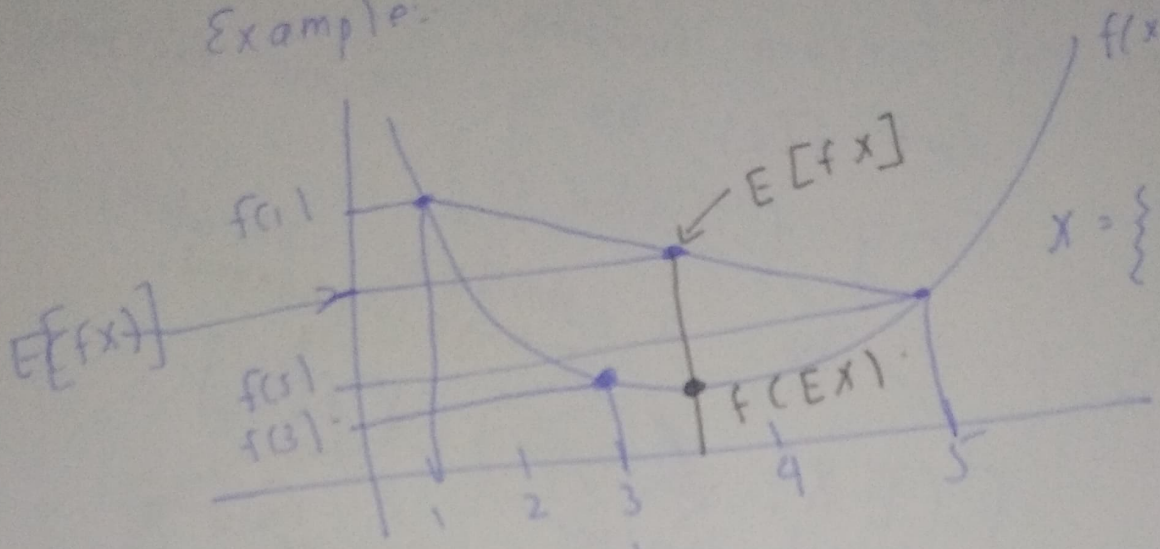
⊙ Let f be a convex function (e.g $f''(x) > 0$)



→ Let X be a random variable

(*) Then $f(EX) \leq E[f(X)]$

Example:



$f(3) = f(EX)$

$E[f(X)] = \frac{1}{2} f(1) + \frac{1}{2} f(5)$

then

$$f(EX) \leq E[f(X)]$$

Further, if $f''(x) > 0$ (f is strictly convex)

then:

$$E[f(X)] = f(EX) \iff X \text{ is a constant.}$$

$X = E[X]$ with probability 1

⇝ Form of Jensen inequality we are going to use for EM algorithm is for concave function:

let $f(x)$ be a concave function:-

(e.g $f''(x) < 0$)

let X be a random variable:-

→ sign is changed

$$f(EX) \leq E[f(X)]$$

(Round the in.

⇝ Density Estimation problem:-

→ Have model for $P(x, z; \theta)$.

→ You only observe X

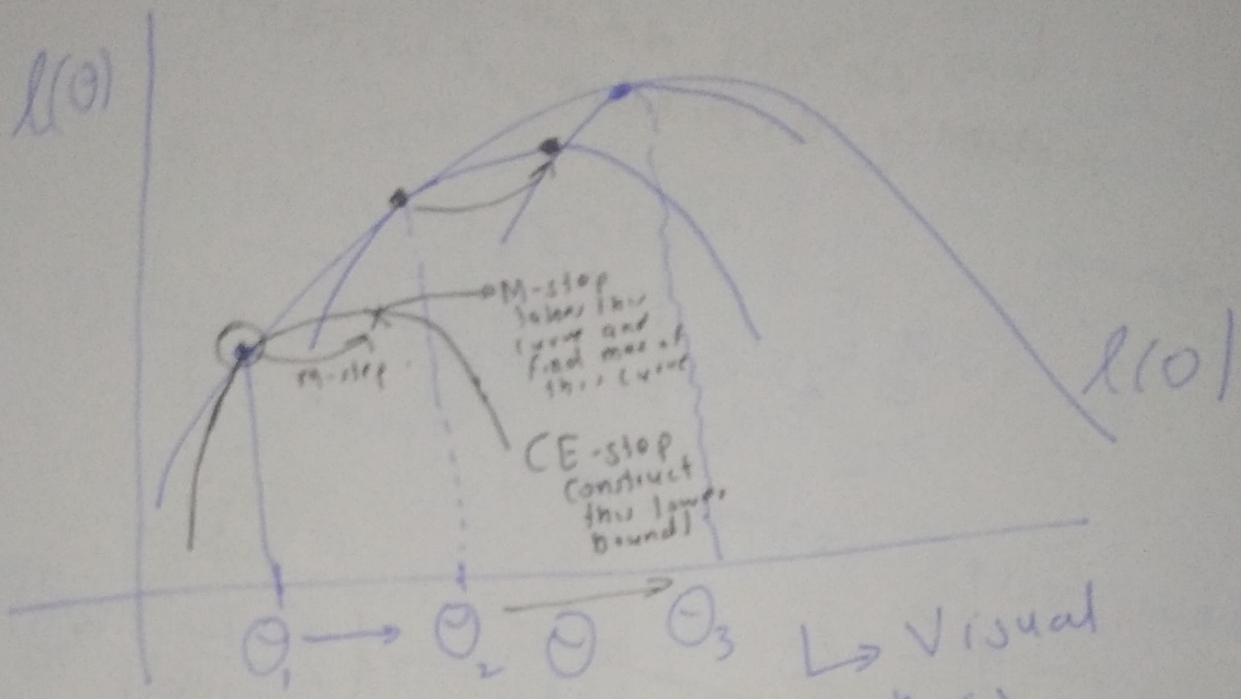
$\{x^{(1)}, \dots x^{(m)}\}$.

~~$\ell(\theta) = \sum_{i=1}^{m} \log(P(x^{(i)}; z^{(i)}; \theta))$~~

$$\ell(\theta) = \sum_{i=1}^{m} \log P(x^{(i)}; \theta).$$

$$\ell(\theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta).$$

Want: $\arg\max_\theta \ell(\theta)$



↳ Visual

Representation of E-step which is construct the lower-bound curve and M-step taking the max of that curve.

$$\max_\theta \sum_i \log p(x^{(i)};\theta)$$

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

$$= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \left[\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right]$$

= where $Q_i(z^{(i)})$ is a probability distribution i.e $\sum_{z^{(i)}} Q_i(z^{(i)}) = 1$.

↳ We will decide later the probability distribution of $Q_i(z^{(i)})$.

Continuing.

$$= \sum_i \log E_{z^{(i)} \sim Q_i} \left[\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right]$$

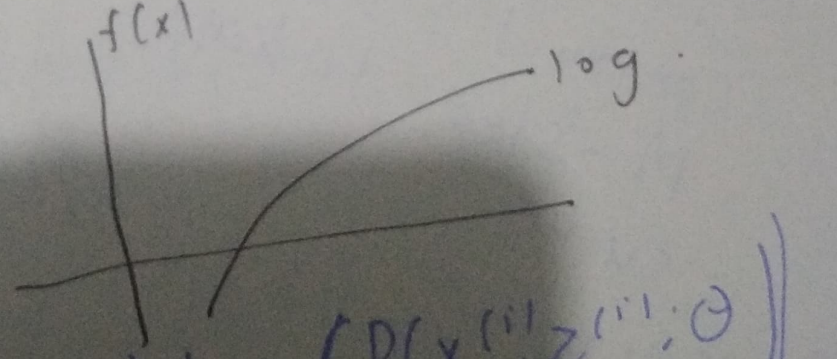
Now using concave form of Jenson inequality :(

$$\sum_i \log \underset{z^{(i)} \sim Q_i}{E} \left[\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right] \geq \sum_i \underset{z^{(i)} \sim Q_i}{E} \left[\log \left(\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right) \right]$$

② ④

$f(E(x))$ $\quad$ $E(fx)$

where $f(x) = \log(x)$



$$① = \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \left(\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right)$$

[Unpacking the terms]

(*) The above function (1) is just a function of θ where:-

→ Function (1) is lower bound (See the last graph) for log likelihood

→ Function (2) is log likelihood.

(*) We have (2) >/ (1) but we want (2) = (1) as shown in graph for current value of theta.

(*) On a given iteration of EM w.r.t parameters θ we want want:-

$$\log E_{z^{(i)} \sim Q_i}\left[\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right] = E_{z^{(i)} \sim Q_i}\left[\frac{\log(P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right.$$

→ For this to be true (remember Jensen inequality) we want:-

$$\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = \text{constant}$$

(Ratio of denominator and numerator must be same).

Therefore Set $Q_i(z^{(i)}) \propto P(x^{(i)}, z^{(i)}; \theta)$

↓ Proportional.

(144)

(157)

(8) To ensure positionality

~~$Q_i(z^{(i)}) = P(x^{(i)}, z^{(i)}$~~

$$\sum_{z^{(i)}} Q_i(z^{(i)}) = 1$$

$$Q_i(z^{(i)}) = \frac{P(x^{(i)}, z^{(i)}; \theta)}{\sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta)}$$

... (Couple More Steps)

$$Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}; \theta)$$

↳ This will ensure the equality of expression (1) and (2).

→ Giving a Summary of EM algorithm:-

E-step:-

Set $Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}; \theta)$ ← Previously $y^{(i)}$

M-step

$$\theta := \arg\max_{\theta} \sum_i P(x^{(i)}$$

$$= \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \left(\frac{P(x^{(i)}, z^{(i)}; \theta}{Q_i(z^{(i)}} \right.$$