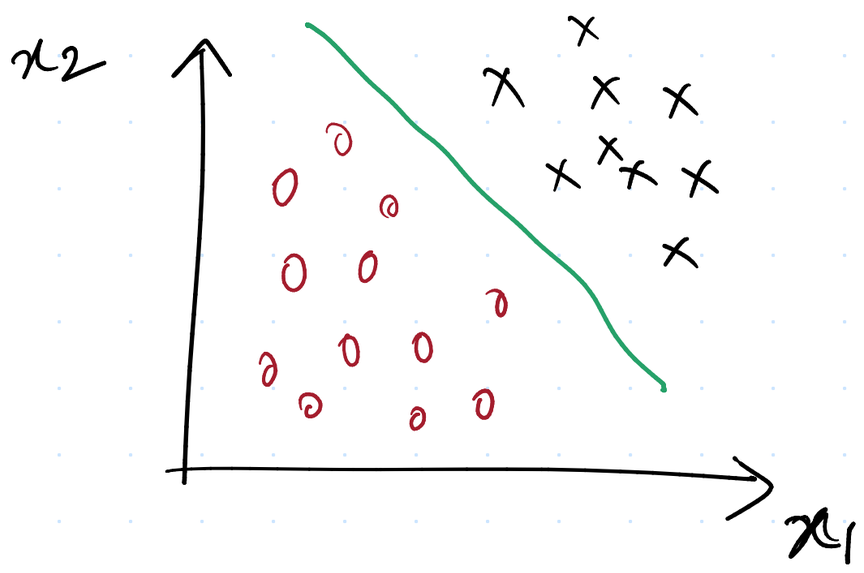
08/03/25

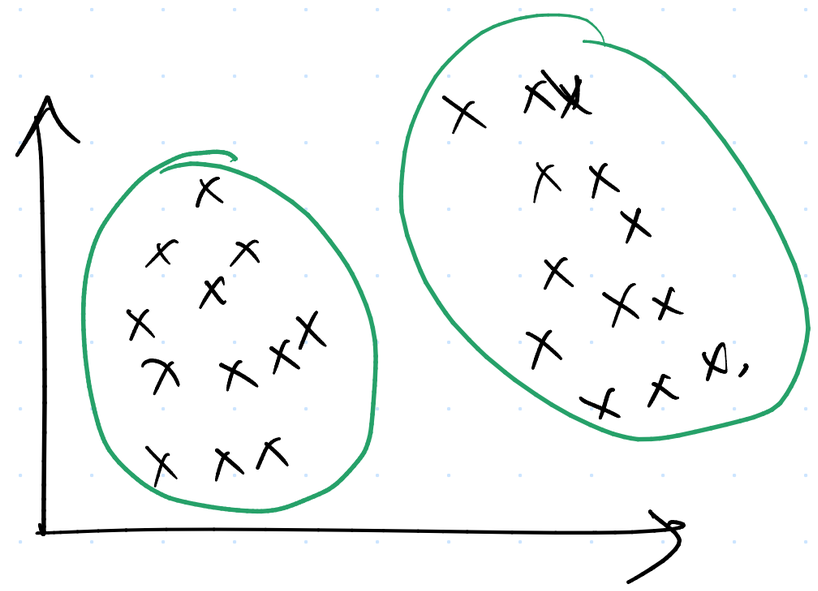
Unsupervised Learning, GMM, EM & VAEs using EMs.

Unsupervised Learning

$X \not\to Y$ (Supervised learning)

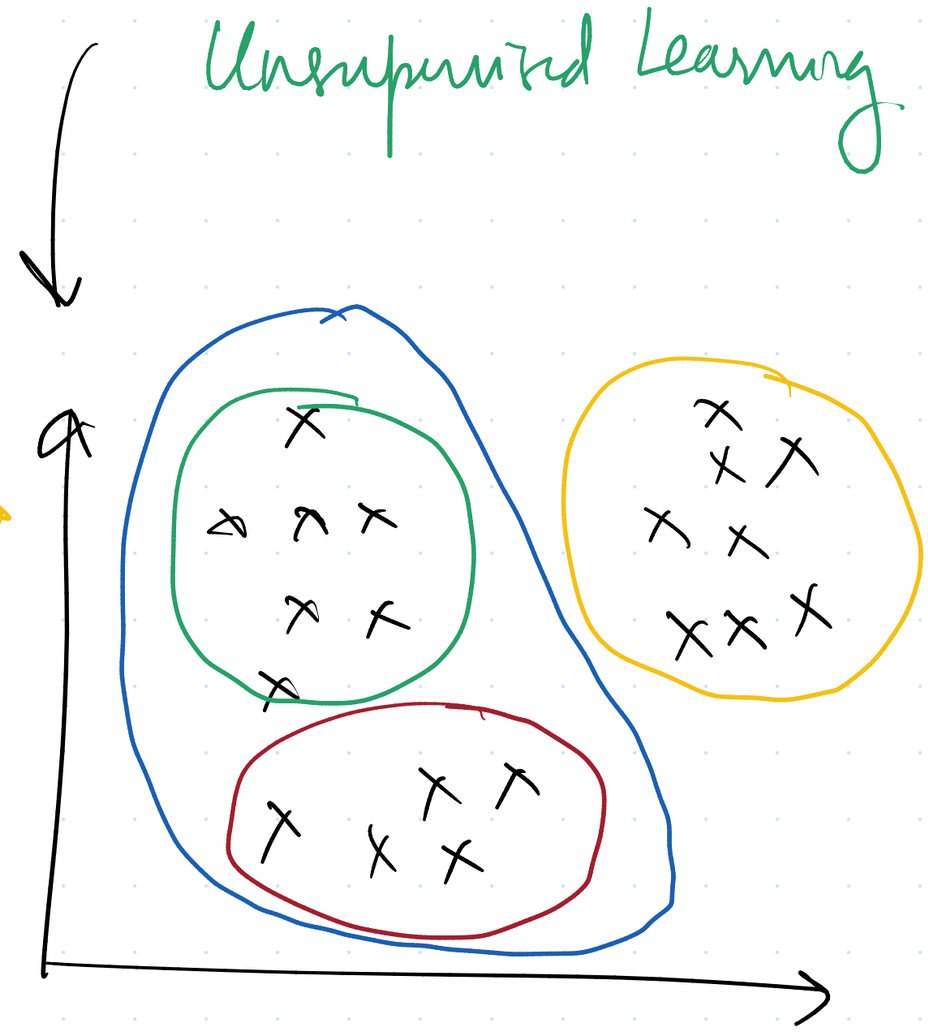
$S = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$.





Supervised Learning.

Unsupervised Learning



Unsupervised Learning → [3]

→ [2]

No correct way to think that only a particular cluster assignment is possible. & there will always be "ambiguity" here

Classification problems in supervised settings can be related to clustering problems in unsupervised (settings) learning.

K-Means clustering Algorithm

$S = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, $x^{(i)} \in \mathbb{R}^d$

K-clusters are given.

1. Initialize the cluster centroids :-

$\mu_1, \mu_2, \mu_3, \ldots, \mu_k \in \mathbb{R}^d$ → randomly.

μ_i's → vector

2. Repeat until convergence

- For every i, set,

$$c^{(i)} = \arg\min \|x^{(i)} - \mu_j\|_2^2$$

- For every j, set :-

$$\mu_j = \frac{\sum_{i=1}^{n} \mathbb{1}\{c^{(i)} = j\}\, x^{(i)}}{\sum_{i=1}^{n} \mathbb{1}\{c^{(i)} = j\}}.$$

mean of all x_i's for which the $(c^{(i)} = j)$

Distortion function.

$$J(c,\mu) = \sum_{i=1}^{n} \| x^{(i)} - \mu_{c^{(i)}} \|_2^2$$

K-means → minimizing the loss of the distortion function in the form of (coordinate descent) → variation of gradient descent.

Instead of minimizing the loss w.r.t. all the variables, we minimize the loss w.r.t. few variables by keeping the others fixed.

Step → 1 : μ - fixed & optimize c

Step → 2 : c - fixed & optimize w.r.t. μ.

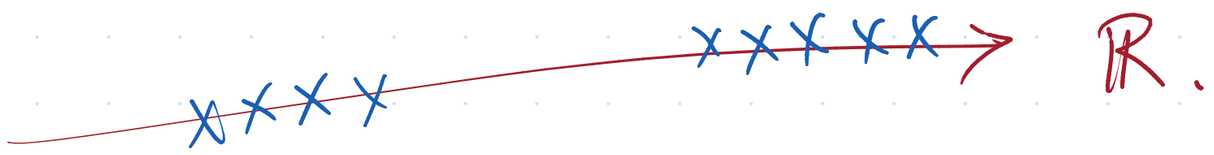
K-means converge eventually since we will reach some form of local minima of the J-function. We may toggle b/w two sets of μ's & c's alternatively once we reach a local minima.

J - non-convex, μ's & c's that we end up in can change from run to run. If we start with a different

initialization, we might end up with different μs & Cs.

(Non-convex problem).

Density Estimation

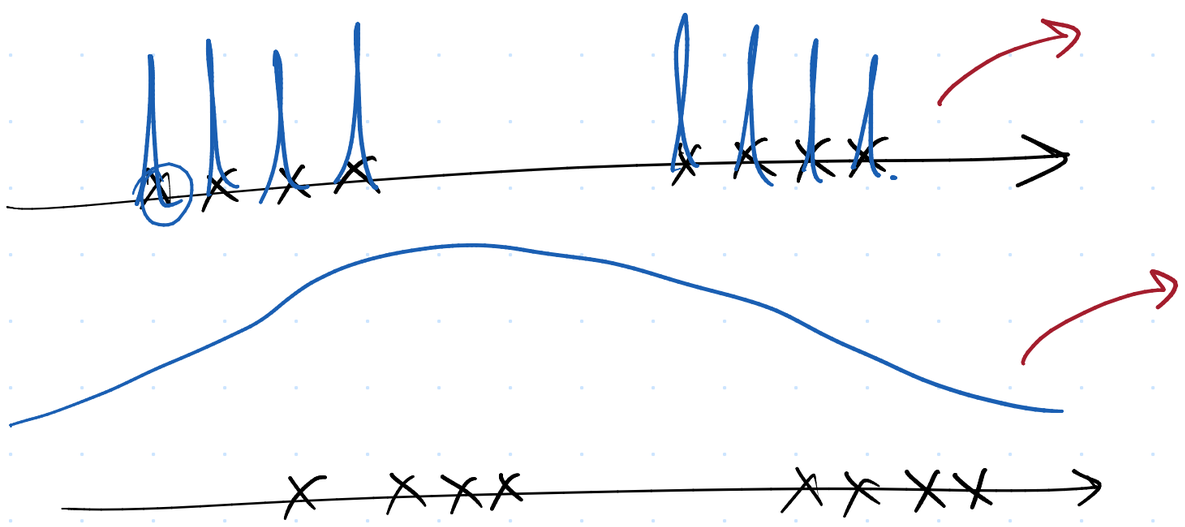


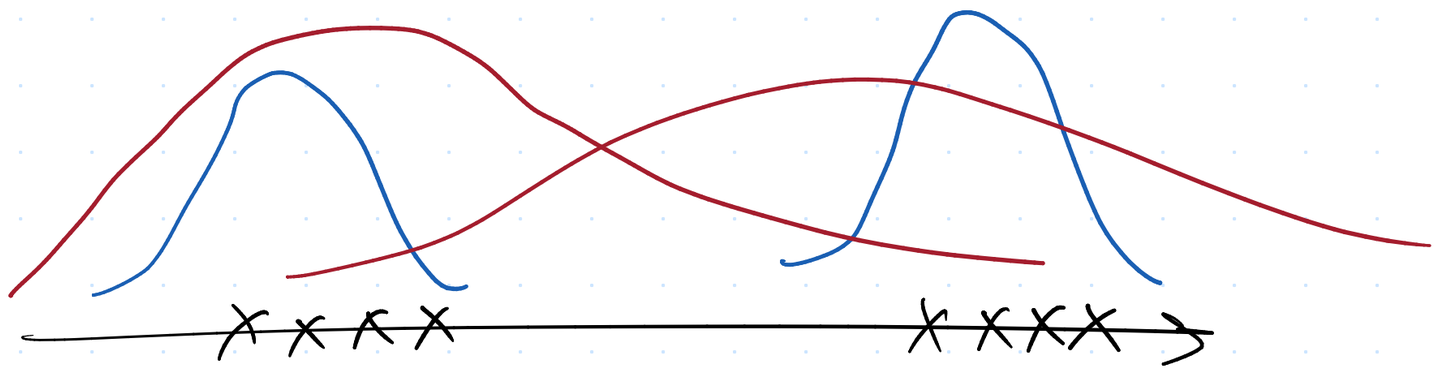
Points are residing in a continuous space. Sampled from some kind of probability distribution. The points come from a continuous distribution, the corresponding prob. dist has some kind of density.

PMF x PDF ✓

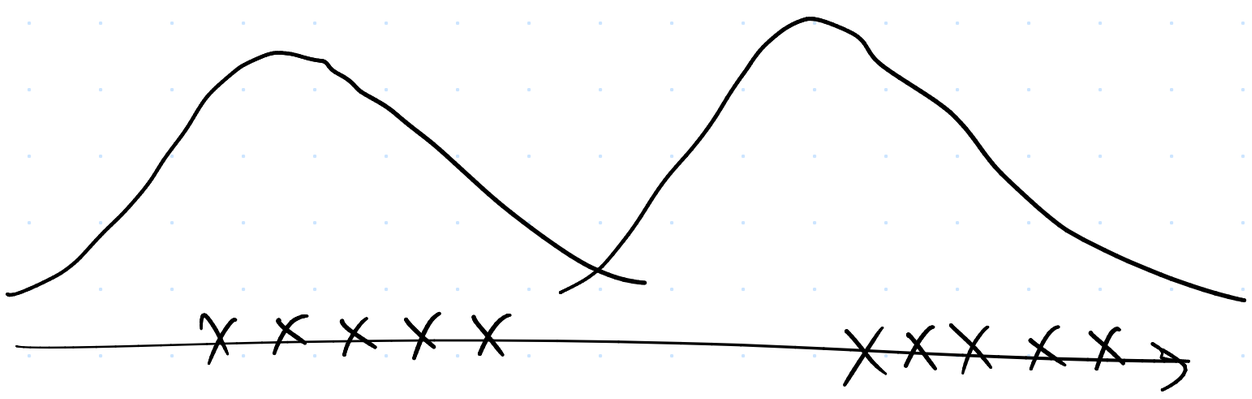
Not a probability mass function, but a probability density function.

Hard problem



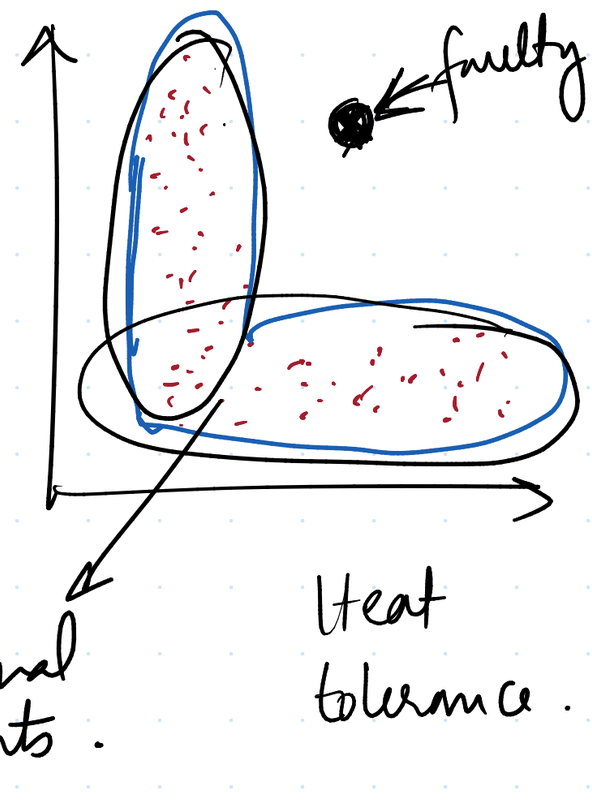


Gaussian Mixture Model :- Mixture of Gaussians.



$p(x) \rightarrow$ to be observed values.

mixture of Gaussians $\longrightarrow$



(E.M. - Algorithm)

$S = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$.

$z^{(i)} \sim$ Multinomial (ϕ)

$\phi_j \geq 0 \qquad \sum_{j=1}^{k} \phi_j = 1$

$\phi_j = p(z^{(i)} = j)$

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

$z^{(i)} \rightarrow$ latent variable, fancy name for R.V. that has not been observed.

GDA : Maximum Likelihood Estimation in GDA.

$$\log p(x, y; \mu, \Sigma, \phi) = l(\mu, \Sigma, \phi)$$

GMM

$$\boxed{\log p(x; \mu, \Sigma, \phi)} = l(\mu, \Sigma, \phi)$$

↑ maximize

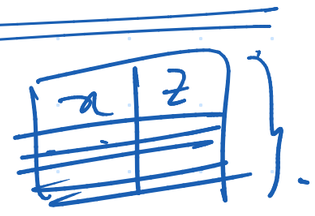
$$= \boxed{\log \sum_z p(x, z; \phi, \mu, \Sigma).}$$

Writing out the full joint dist & marginalizing out the latent variable.

If we have observed z, then it could have contributed to the likelihood objective, but we don't know what z is, so it cannot be written - formulated like that.

$p(z)$ → class prior / if continuous.

$p(x, z)$ → model



z → latent (unobserved)

$p(z|x)$ = posterior

$p(x|z) = \frac{p(z|x)\,p(x)}{p(z)}$ ← discriminating process.

↑ Generating process.

$p(x)$ → evidence.

Goal = maximize the likelihood given the evidence.

EM Algorithm — inspired by K-Means.

Repeat until convergence (randomly initialize μ, ϕ, Σ)

given a point, we assign a prob. that this point belongs to a particular centroid.

E-Step

For each i, j, set :- $w_j^{(i)} := p(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma)$

weight / posterior dist.

M-Step

Update the parameters

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

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

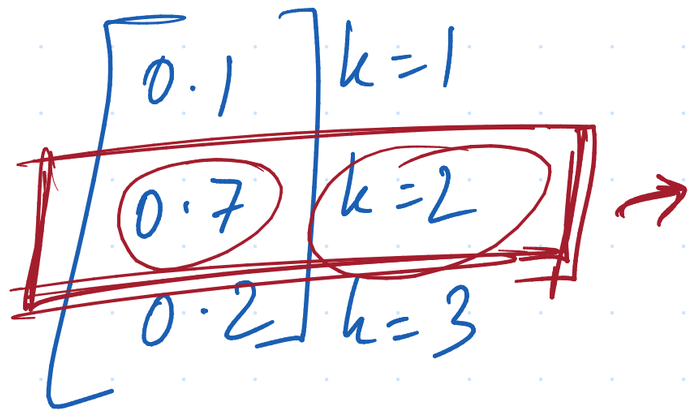
$$\Sigma_j = \frac{\sum_{i=1}^{n} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)}}$$

We don't need to know $z^{(i)}$, we are constructing the probability $z^{(i)} = j$ using the Bayes rule.

(Soft k-means)

$k = 3$,

$p(z^{(i)} = j \mid x^{(i)})$ ⇒



We are doing soft assignment, where every point has a soft assignment in the form of probability distribution.

k-means → hard assignment → where each point belongs to one cluster only.

$$p(z|x) = \frac{\overbrace{p(x|z)}^{\text{Gaussian}} \overbrace{p(z)}^{\text{Multinomial}}}{\sum_z p(x|z)\ p(z).}$$

→ this takes the form of a softmax.

EM algorithm gives a framework for performing MLE when some variables are unobserved.

Used in cases where we have a functional form of $p(x, z)$. x & z → could be anything, but z is unobserved.

Expectation Maximization

MLE in the presence of latent variables.

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

If x & z → observed, we perform simple MLE.

But when z → unobserved, we instead perform maximization.

$l(\theta) = \log p(x; \theta)$

EM gives us a framework for achieving this.

Jensen's inequality → probabilistic tool.

f → to be convex function

$f''(x) \geq 0 \; \forall x.$

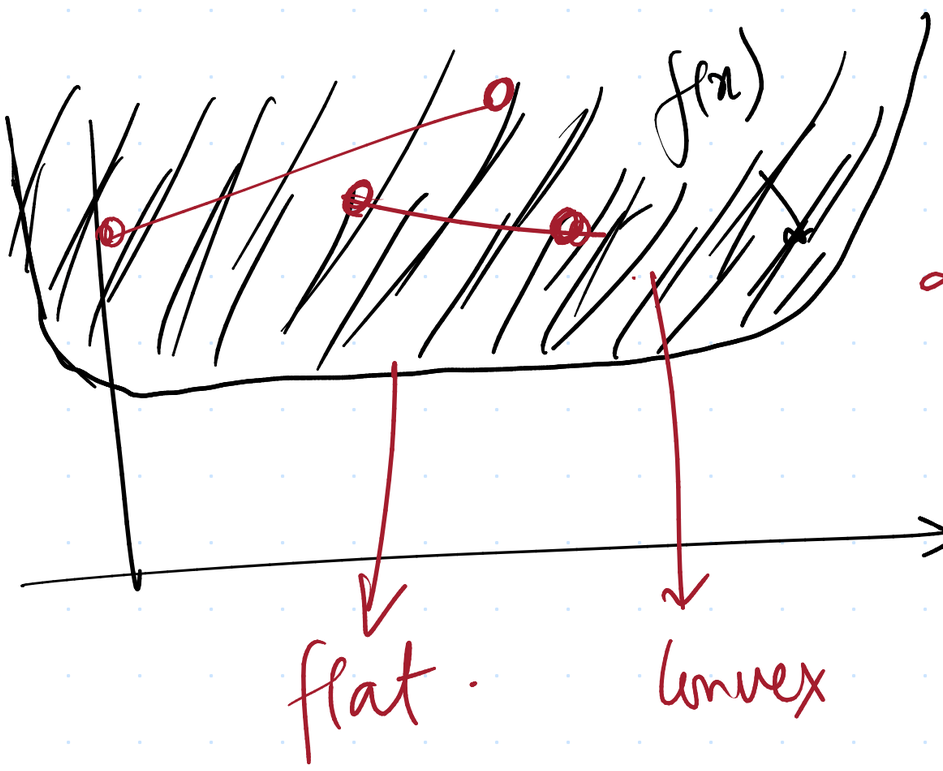
f → is strictly convex if, $f''(x) > 0 \; \forall x.$

$E[f(x)] \geq f(E[x])$ → where f is Convex.

↓ R.V.

Convex function

(⚠ How to define this region?)



$f(x) = 2x^2$

$f'(x) = 4x$

$f''(x) = 4$

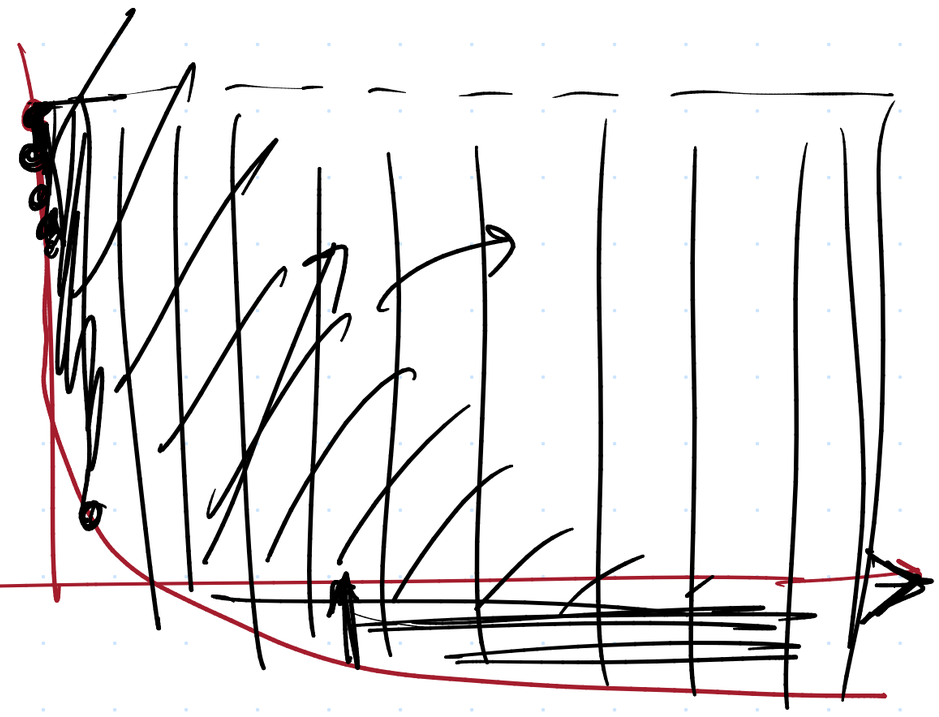
$f''(x) \geq 0$?

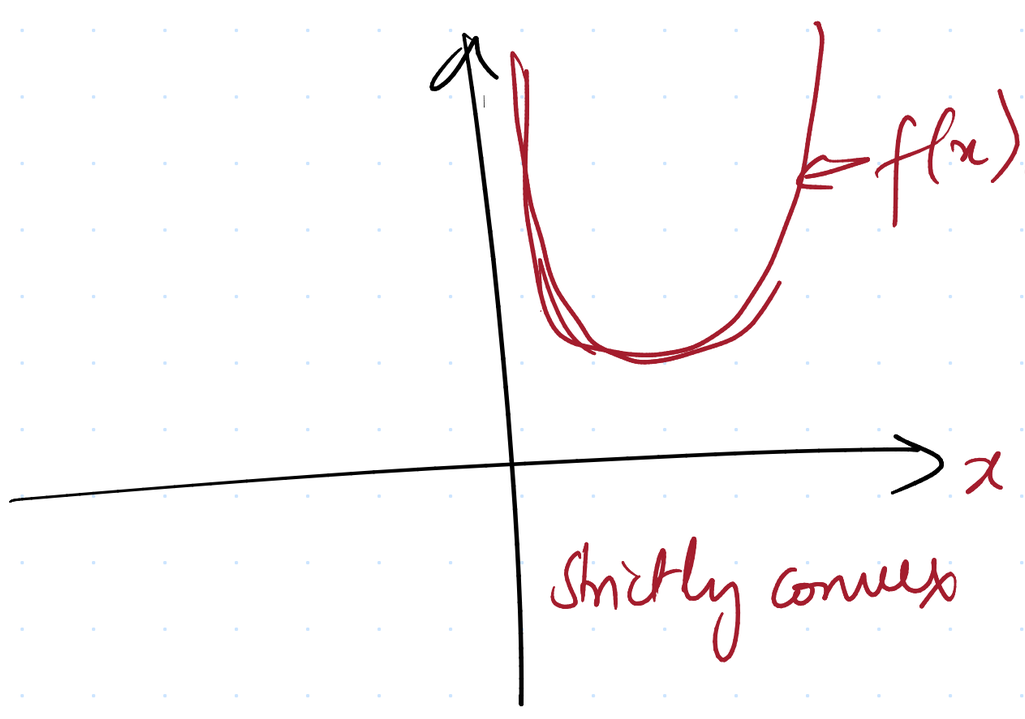
Convex.

If f is strictly convex, then,

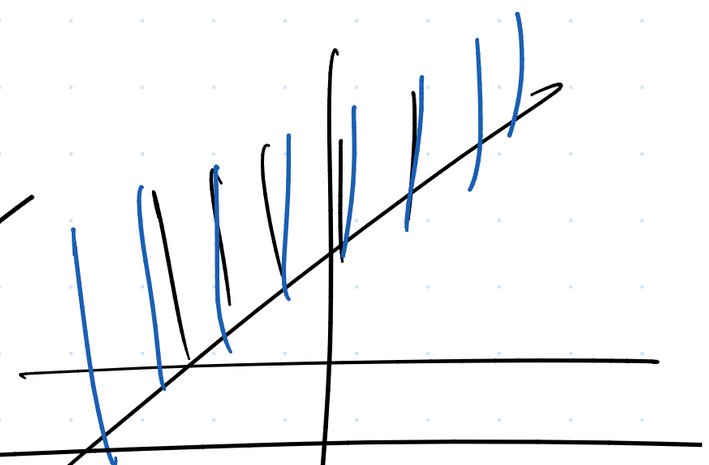
If $E[f(a)] = f(E[x])$ then,

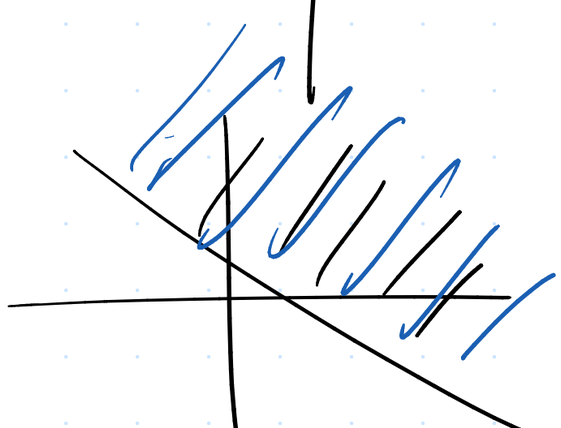
$x = E[x]$, with prob = 1.





$f''(x) \geq 0$ convex	$f''(x) \leq 0$ concave	strict. $>$ either or $<$
$f(x) = x^2$ ✓	$f(x) = -x^2$ ✓	✓
$f(x) = \ln x + C$. ✓	✓	✗
$e^x \rightarrow$ convex	$-e^x \rightarrow$ concave	✓
$e^{-x} \rightarrow$	$-e^{-x} \rightarrow$	✓
$-\log(x) \rightarrow$ convex	$\log(x) \rightarrow$ concave.	✓



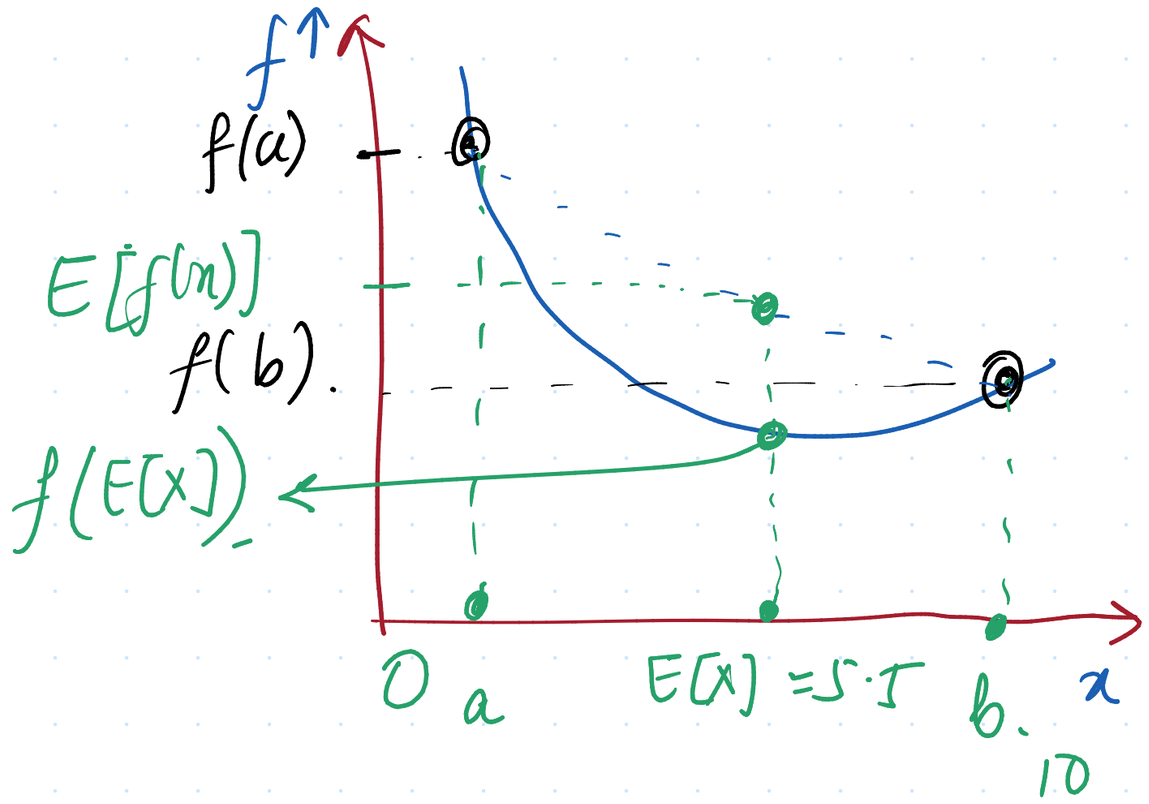


✓ If f is concave, e.g. $f(x) = \log(x)$

$$E[f(x)] \leq f(E[x])$$

Jensen's Equality

$$E[\log(x)] \leq \log(E[x])$$



E.M. Algorithm

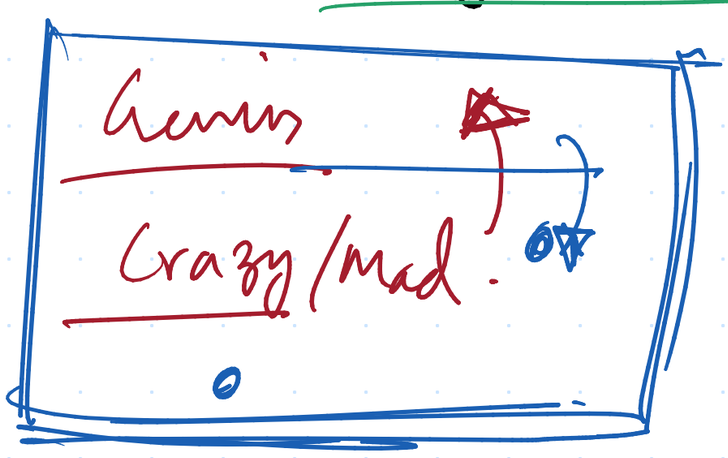
Goal: Maximize $\sum_{i=1}^{n} \log p(x^{(i)}; \theta)$.

z's were unobserved, hence maximizing this is very hard.

$\max \log p(x;\theta)$

Summation over z, here z will be continuous and hence the integration will be arbitrarily complex & hard.

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



$$= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)}$$

$Q(z)$ ← probability. $Q(z) > 0 \ \forall z$.

$Q(z)$ ↑ function.

$$= \log \mathop{\mathbb{E}}_{z\sim Q(z)} \left[\frac{p(x,z;\theta)}{Q(z)} \right]$$

Applying Jensen's Equality →

$$\geq \mathop{\mathbb{E}}_{z\sim Q(z)} \left[\log \frac{p(x;z;\theta)}{Q(z)} \right]$$

log → concave function.

ELBO
Evidence Lower Bound.

$:= ELBO(x; Q, \theta)$

If we are finding values of θ & ϕ, then the same values of θ, implicitly the $\log p(x;\theta)$ is going up.

Since, $ELBO(x;\phi,\theta) \leq \log p(x;\theta)$ hence this is true.

There are cases when,

$$\boxed{\log p(x;\theta) = ELBO(x;\phi,\theta)}$$

$$\log \mathop{\mathbb{E}}_{z\sim q}\left[\frac{p(x,z;\theta)}{q(z)}\right] = \mathop{\mathbb{E}}_{z\sim q} \log\left[\frac{p(x,z;\theta)}{q(z)}\right]$$

↓

if this is a constant.

Since, log is strictly convex :-

$\frac{p(x,z;\theta)}{q(z)}$ → constant, w.r.t z. $= c$.

$$q(z) = \frac{1}{c}\, p(x,z;\theta)$$

→ proportionality constant.

$$Q(z) \propto p(x, z; \theta)$$

$$Q(z) = \frac{p(x, z; \theta)}{\sum_z p(x, z; \theta)}$$

distribution over z that must sum to 1.

no normalization constant similar to the proportionality constant.

$$= \frac{p(x, z; \theta)}{p(x, \theta)} = p(z|x; \theta).$$

$\sum Q(z) = 1$.

x & z → doesn't have any specific form & it can be mixture of Gaussians, it could be any algorithm etc.

This holds for any latent variable models, hence, we will apply it to optimize / train VAE's in the next lecture.

So, EM Algorithm :-

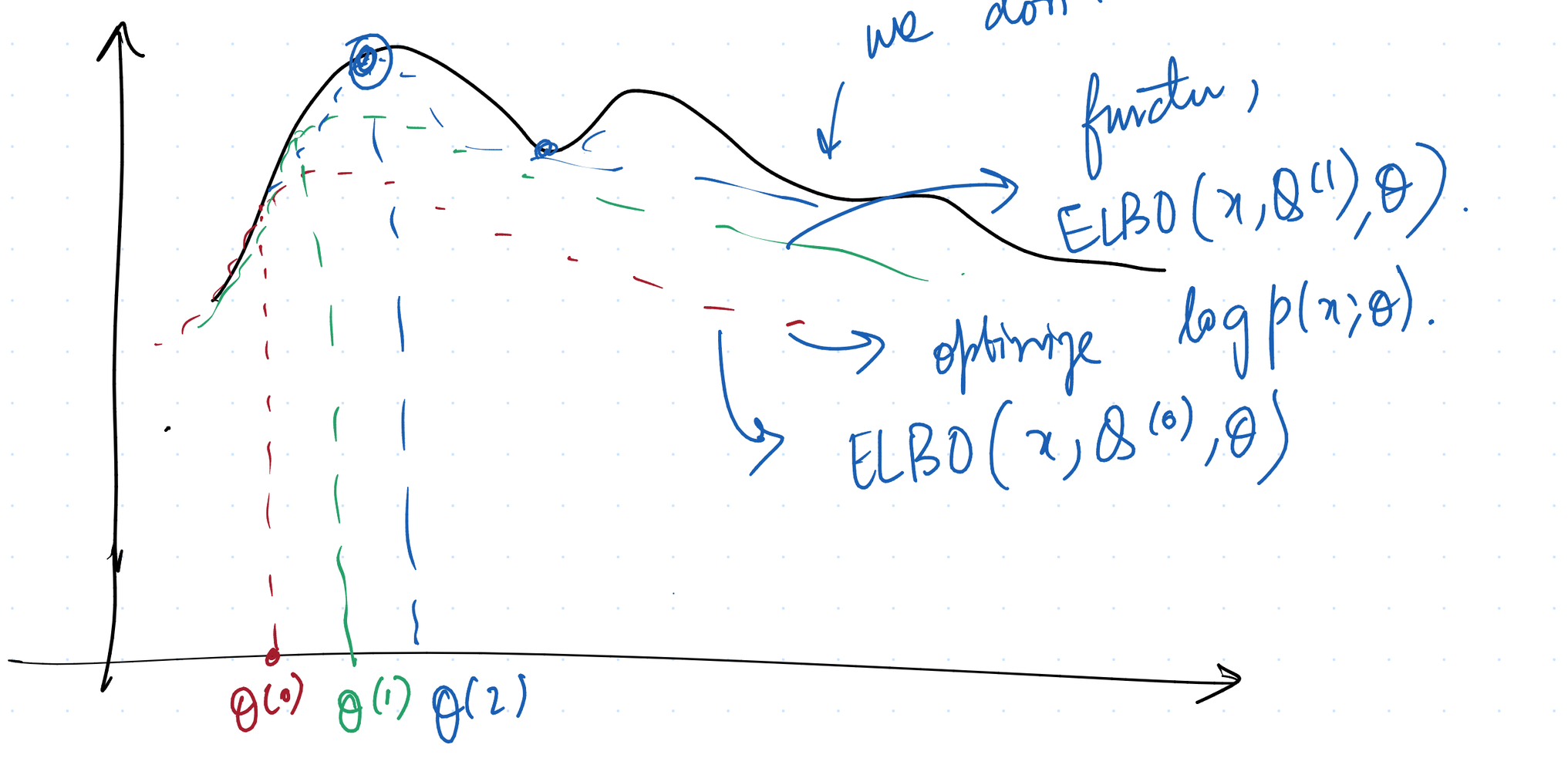
E-step

For each i, set :-

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

M-step

$$\theta := \arg\max_{\theta} \sum_{i=1}^{N} ELBO(x^{(i)}; Q_i; \theta).$$



So, in every step, the ELBO that we get is tight & will be touching the $\log f(x)$.