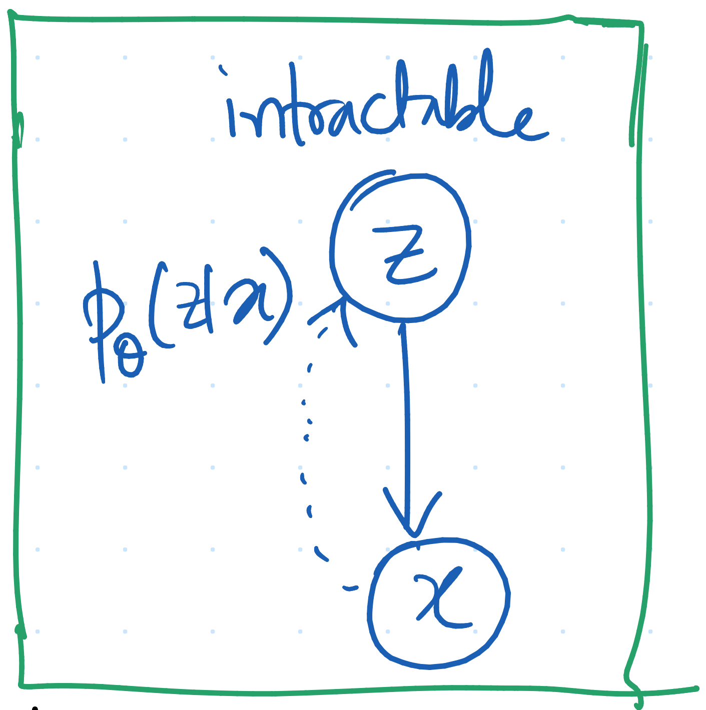
VAE using Expectation Maximization strategy. 16/3/25

M.C.M.C. Expectation Maximization.

→ MCMC sampling.



Assumption :-

We can calculate $p(z|x^{(i)})$ → this maynot hold in practice.

E-step : For all i

set $Q_i^{(t)}(z) = p(z \mid x^{(i)}; \theta^{(t)})$

M-step

$$\theta^{(t+1)} = \arg\max \sum_{i=1}^{n} \sum_{z} Q_i^{(t)}(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}, \theta)}{Q_i^{(t)}(z)}$$

doesn't depend on θ.

$$= \arg\max_{\theta} \sum_{i=1}^{n} \sum_{z} Q_i^{(t)}(z^{(i)}) \log p(x^{(i)}, z^{(i)}; \theta)$$

(same as maximizing this quantity).

$z \sim N(0, I_{k\times k})$

$x|z \sim \text{NeuralNet}_\theta(z^{(i)}) \sim \text{Complex}(z^{(i)}, \theta)$

$p(z|x)$ → impossible, cannot be calculated, very complex → intractable.

$$\log\left(\frac{A}{B}\right) = \log A - \cancel{\log B} \rightarrow \text{constant}$$

w.r.t. θ

hence, we get rid of this term.

$$\Rightarrow \arg\max_{\theta} \sum_{i=1}^{n} \sum_{z} Q_i^{(t)}(z^{(i)}) \log p(x^{(i)}, z^{(i)}; \theta).$$

$$\Rightarrow \arg\max_{\theta} \sum_{i=1}^{n} \underset{\underline{z^{(i)} \sim Q_i^{(t)}}}{\mathbb{E}} \left[\log p(x^{(i)}, z^{(i)}; \theta)\right]$$

$Q_i^{(t)}(z)$ in E-step is only used to calculate this expectation.

We are not interested in the density of Q. We are only interested in the density value of Q to perform this expectation.

$$\simeq \arg\max_{\theta} \sum_{i=1}^{n} \frac{1}{M} \sum_{m=1}^{M} \log p(x^{(i)}, z^{(m)}; \theta)$$

$$\boxed{z_i^{(m)} \sim Q_i^{(t)}}$$

(replace with Monte-Carlo estimation of the estimation)

$z_i^{(m)}$ = sampled from $Q_i^{(t)}$.

Q- posterior

Sample from posterior of a complex probability distribution even though we don't know how to evaluate it.

Law of large numbers tells us that as the M goes to ∞ the Monte-Carlo estimate will tend to converge towards the true expectation.

{ Gibbs Sampling
Metropolis Hastings. }

We don't know the posterior, but we can approximate the posterior using the Monte-Carlo Technique / Sampling Technique.

Here the guarantee of the likelihood increase after every step doesn't hold anymore, because this is an approximation of the lower bound and not the exact lower bound.

So, there are three ways to do this :-

* Exact posterior → maths / calculation etc. → (diff)
* Gibbs / Sampling methods → ✓.
* Variational Inference → optimization. → (previous class).

Jensen's Inequality.

$$\log p(x) \geq \underbrace{ELBO}(x; \theta)$$

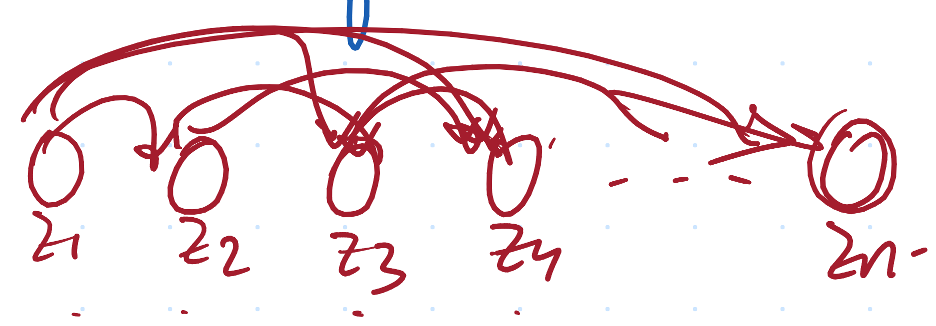
↳ lower bound of the evidence.

$$\log p(x) = \text{ELBO}(x; Q) + \underbrace{\qquad}_{?}$$

$$D_{KL}(Q \| p(z|x)) = \log p(x) - \text{ELBO}(x; Q)$$

$$\underbrace{\log p(x)}_{\substack{\text{const. w.r.t.} \\ Q}} = \underbrace{\text{ELBO}(x; Q)}_{\substack{\text{maximize} \\ \text{w.r.t. } Q.}} + \underbrace{D_{KL}(Q \| p_{z|x})}_{\geq 0.}$$

$$p_{z|x} = \underset{q \in Q}{\operatorname{argmax}} \; \text{ELBO}(x; q) \leftarrow \text{Variational Inference.}$$

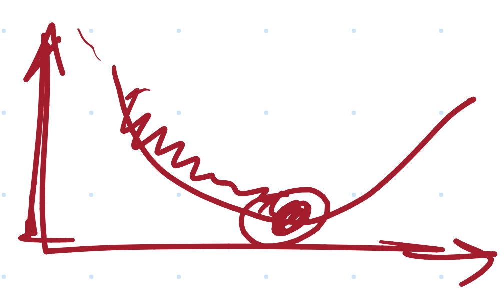


Mean field Assumption:-

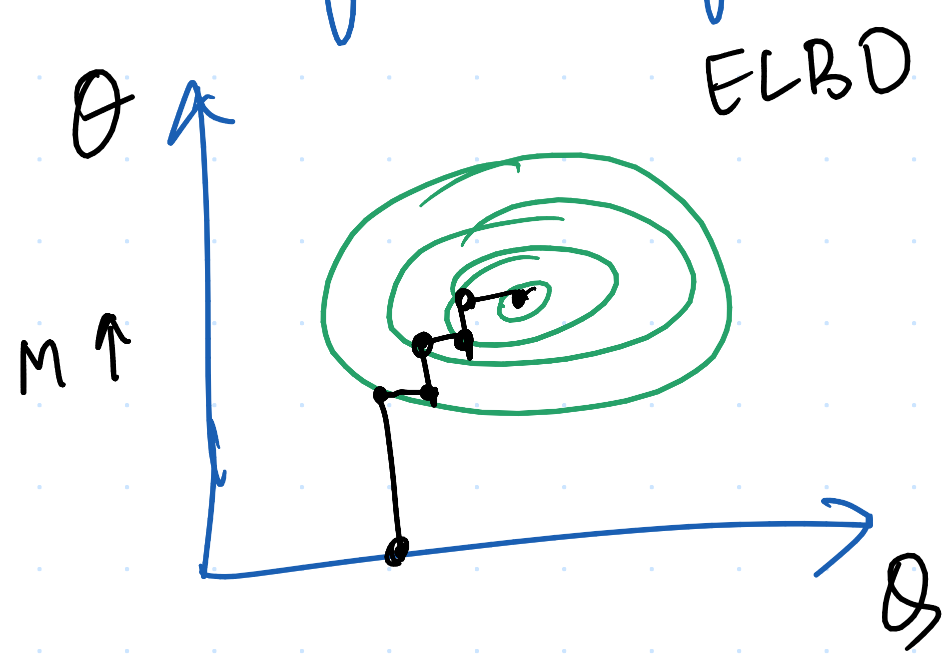
$$Q(z) = Q_1(z_1)\, Q_2(z_2)\, Q_3(z_3) \ldots Q_k(z_k)$$

$$z \in \mathbb{R}^k.$$

If we assume that the component of z vector can be factored into k-independent scalar, probability distribution, then, this is called mean field assumption.



Mean field inference makes computation easier.



co-ordinate ascent - multiple variables. Start with random initialization, all except some fixed & optimize the ones that are not held fixed.

Worked well with classic EM.

Once updated estimates are got, hold them fixed and update the other ones & so on.

Can we do gradient ascent? → Variational A.E.

$z \sim N(0, I_{k \times k})$ → Latent variable / prior → sampled from some normal distribution.

Important.

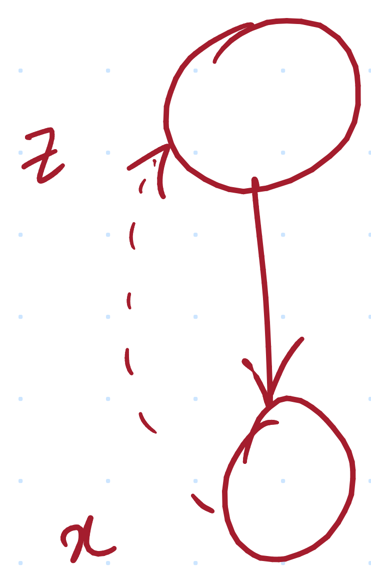
posterior evidence

$$\boxed{p(x|z) = \frac{p(z|x)\, p(x)}{p(z)}}$$

likelihood prior.

$x|z \sim N(g(z;\theta), \sigma^2 I)$
↓
Likelihood.

$p(z|x)$; g → Neural Network with param, θ, we are never going to obtain.



Approximate using variational inference $p(z|x)$

Q → family for Variational Inference (V.I.)

$$\underbrace{Q_i(z)}_{\mathbb{R}^K} = N\Big(\underbrace{q(x^{(i)};\phi)}_{\in \mathbb{R}^K},\ \underbrace{\mathrm{diag}\big(V(x^{(i)};\psi)\big)^2}_{\in \mathbb{R}^{K \times K}}\Big)$$

μ . Σ

We will get different $Q \to$ dist per example. In E.M. E-step is performed separately per example.

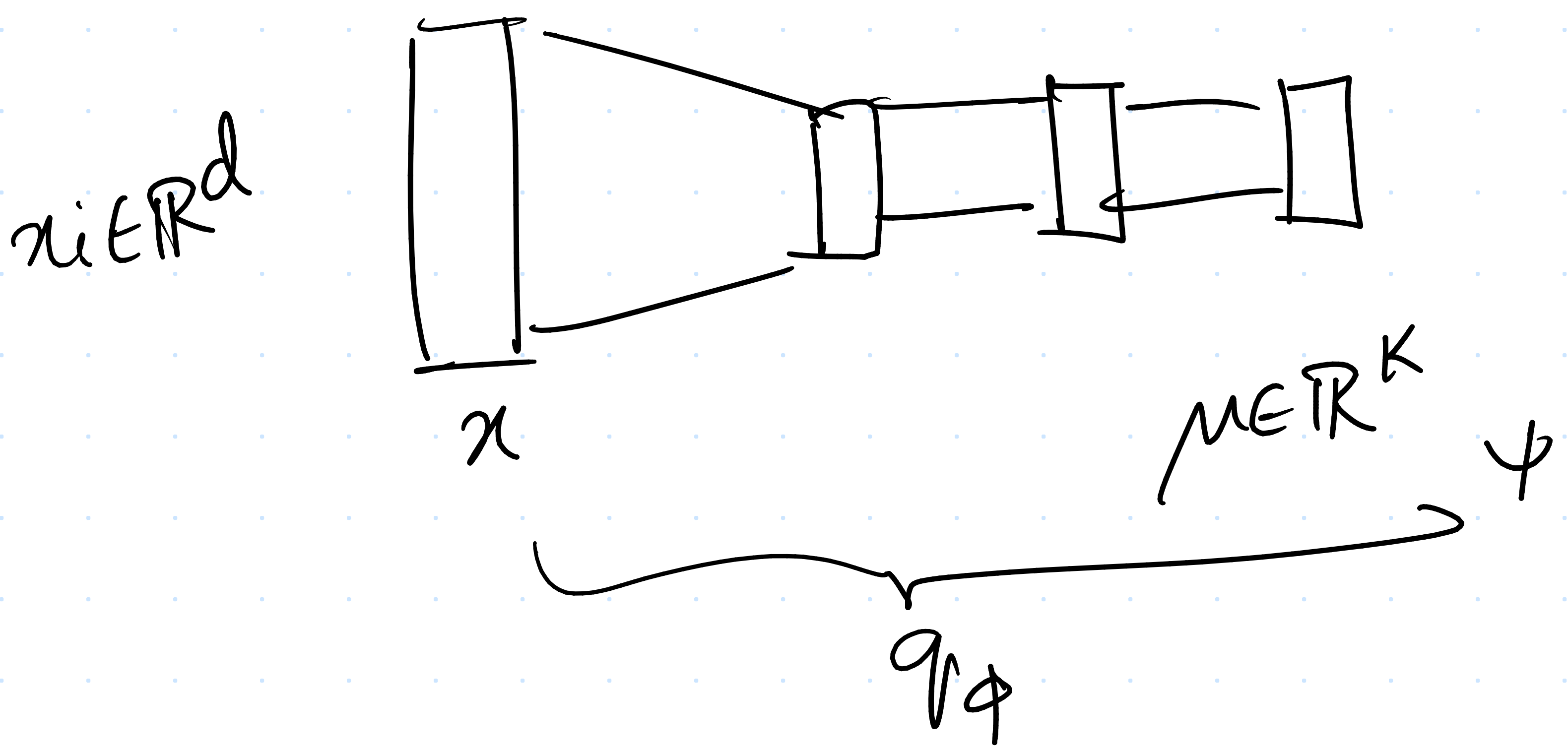
$Q_i \to$ depends on x $\quad Q_i = p(z|x)$ $\quad$ N.N. $\to$ takes x as input

"<u>Amortized Inference</u>":

function of x, and those function is the N.N., so we don't have to optimize for each example independently as the E-M step.

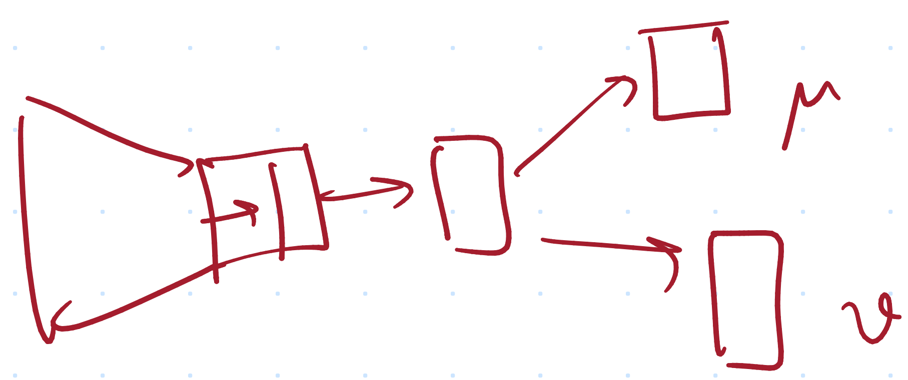
$q_\phi(x^{(i)}) = \mu^{(i)}$

why is $k < d$, $\mu \to$ smaller dim.

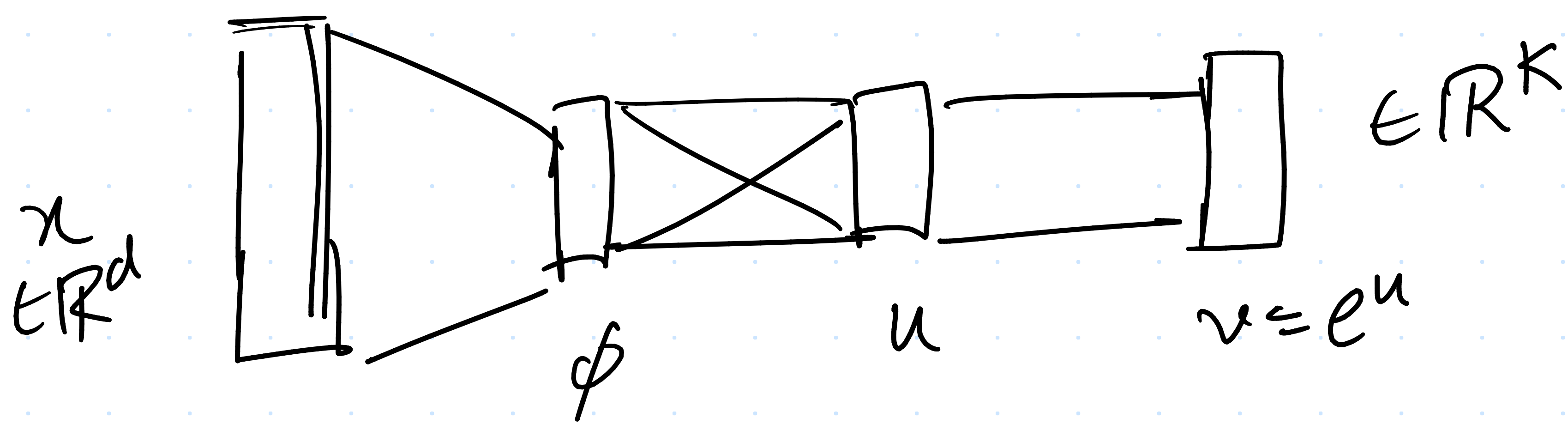


$Q \to$ dist over z.

and $z = k$ dim, $\mu = k$-dim.



<u>Covariance.</u>



to get +ve s.d.

positive = square them, exponentiate each element.

$Q_i \rightarrow$ mean & covariance.

Normal Dist $\rightarrow$ vector & a matrix in high dimensional space.

$$\begin{array}{cc} & k \\ k & \begin{bmatrix} v_1 & & & \\ & v_2 & & \\ & & \ddots & \\ & & & v_k \end{bmatrix} \end{array}$$

Gaussian dist :- & diagonal covariance matrix. If there is no correlation b/w two components of a joint Gaussian, then, they necessarily must be independent.

Diagonal covariance matrix $\rightarrow$ making the mean field assumption. Each of z_i's $\rightarrow$ independent.

ϕ & $\psi \rightarrow$ shared across all the examples.

$\boxed{\text{Decoder } p(x|z) \cdot p(z)}$

$$\text{ELBO}(\phi, \psi, \theta) = \sum_{i=1}^{n} \mathbb{E}_{z^{(i)} \sim Q^{(i)}} \left[\log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right]$$

encoder

where, $Q_i = N(\underbrace{q(x^{(i)}; \phi)}_{\mu}, \underbrace{\text{diag}(v(x^{(i)}, \psi))}_{\Sigma})$.

$$\text{ELBO}(Q, \theta) = \sum_{i=1}^{n} \mathbb{E}_{z^{(i)} \sim Q^{(i)}} \left[\log \frac{P(x, z; \theta)}{Q_i(z^{(i)})} \right]$$

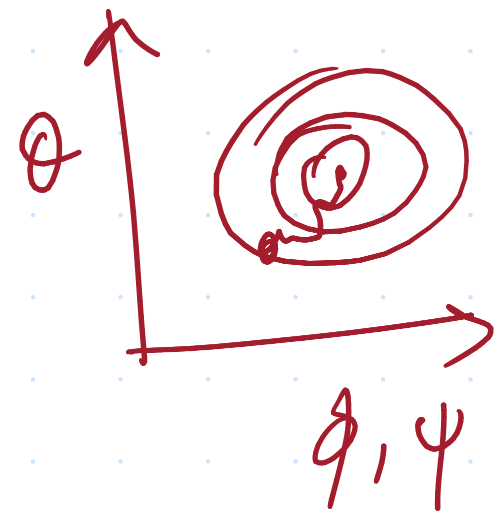
ELBO in EM, where each $Q \to$ separately calculated for each example in parallel.

Optimize ϕ, ψ, θ with gradient descent:-

$$\theta := \theta + \eta \, \nabla_\theta \, \mathrm{ELBO}(\phi, \psi, \theta)$$

$$\phi := \phi + \eta \, \nabla_\phi \, \mathrm{ELBO}(\phi, \psi, \theta)$$

$$\psi := \psi + \eta \, \nabla_\psi \, \mathrm{ELBO}(\phi, \psi, \theta)$$



$$\nabla_\theta \, \mathrm{ELBO}(\phi, \psi, \theta)$$

$$= \nabla_\theta \sum_{i=1}^{N} \mathbb{E}_{z^{(i)} \sim Q_i} \left[\log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right]$$

$$= \sum_{i=1}^{N} \mathbb{E}_{z^{(i)} \sim Q_i} \left[\nabla_\theta \log p(x^{(i)}, z^{(i)}; \theta) \right]$$

$$= \sum_{i=1}^{N} \mathbb{E}_{z^{(i)} \sim Q_i} \left[\underbrace{\nabla_\theta \log p(x^{(i)} | z^{(i)}; \theta)}_{\text{Generator Decoder}} + \cancelto{0}{\nabla_\theta \log p(z^{(i)})} \right]$$

Monte Carlo (M.C.)

$$\nabla_\phi \, \mathrm{ELBO}(\phi,\psi,\theta) = \nabla_\phi \sum_{i=1}^{N} \mathbb{E}_{z^{(i)}\sim Q_i}\left[\log \frac{p(x,z;\theta)}{Q(z^{(i)})}\right]$$

The distribution $z^{(i)} \sim Q_i$ depends on ϕ, hence we cannot swap the expectations.

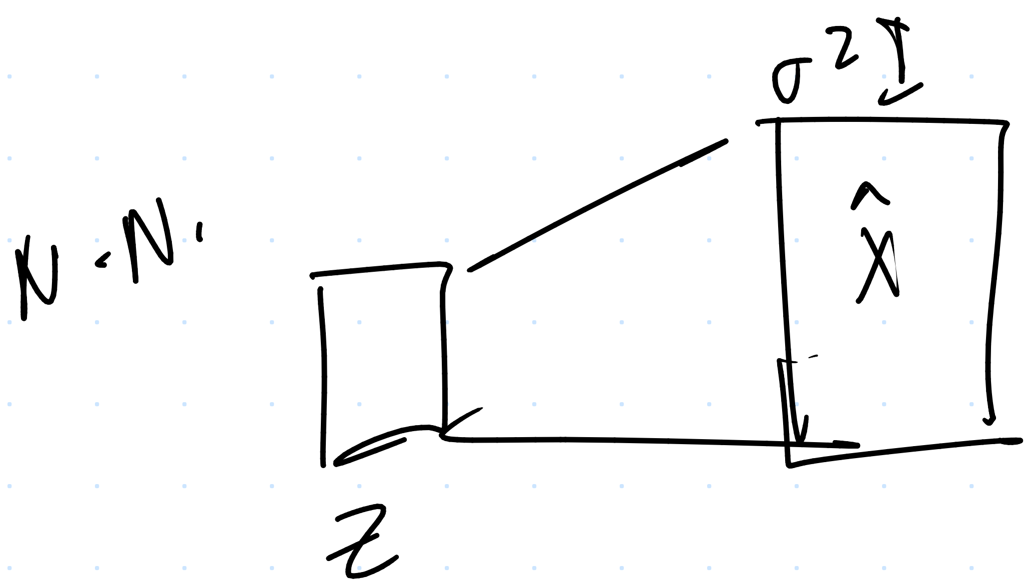
decoupled into this.

$z \sim N(\mu, \sigma)$ $\qquad\qquad$ $\varepsilon \sim N(0,1)$

$$\boxed{z = \varepsilon \cdot \sigma + \mu}$$

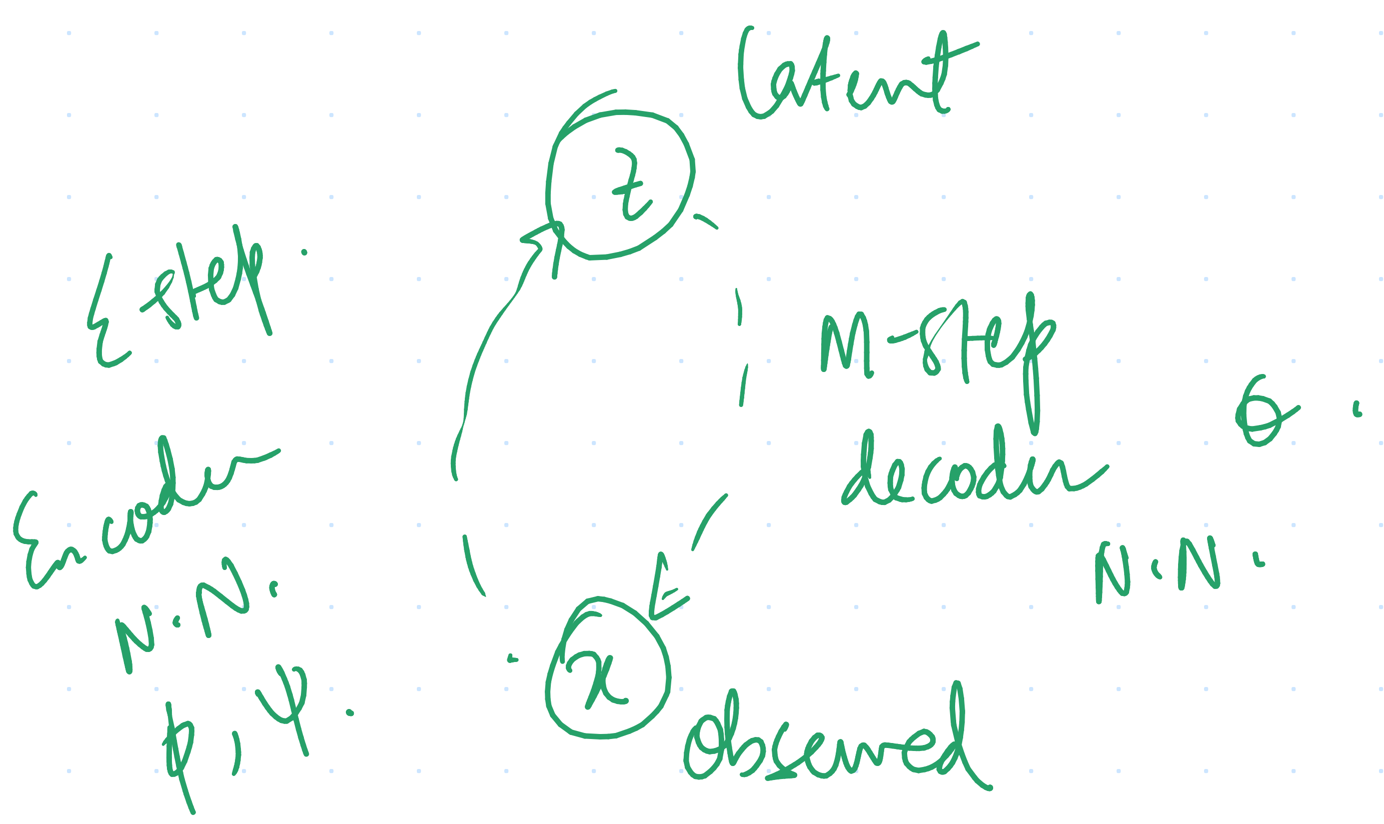
$$\nabla_\phi \sum_{i=1}^{N} \mathbb{E}_{\varepsilon^{(i)}\sim N(0,1)}\left[\log \frac{p(x, \overbrace{\boxed{\varepsilon^{(i)} \odot \Sigma^{(i)} + \mu^{(i)}}}^{z^{(i)}}; \theta)}{Q(\underbrace{\boxed{\varepsilon^{(i)} \odot \Sigma^{(i)} + \mu^{(i)}}}_{z^{(i)}})}\right]$$

where, $\mu^{(i)} = q(x^{(i)};\phi)$ $\qquad$ $\Sigma^{(i)} = \mathrm{diag}(v(x^{(i)});\psi)$.



$= \mu$ $\qquad$ $p(x^{(i)} | z^{(i)}; \theta)$.

$$\log p(x^{(i)}; g(z^{(i)}), \sigma^2 I)$$



$\downarrow$

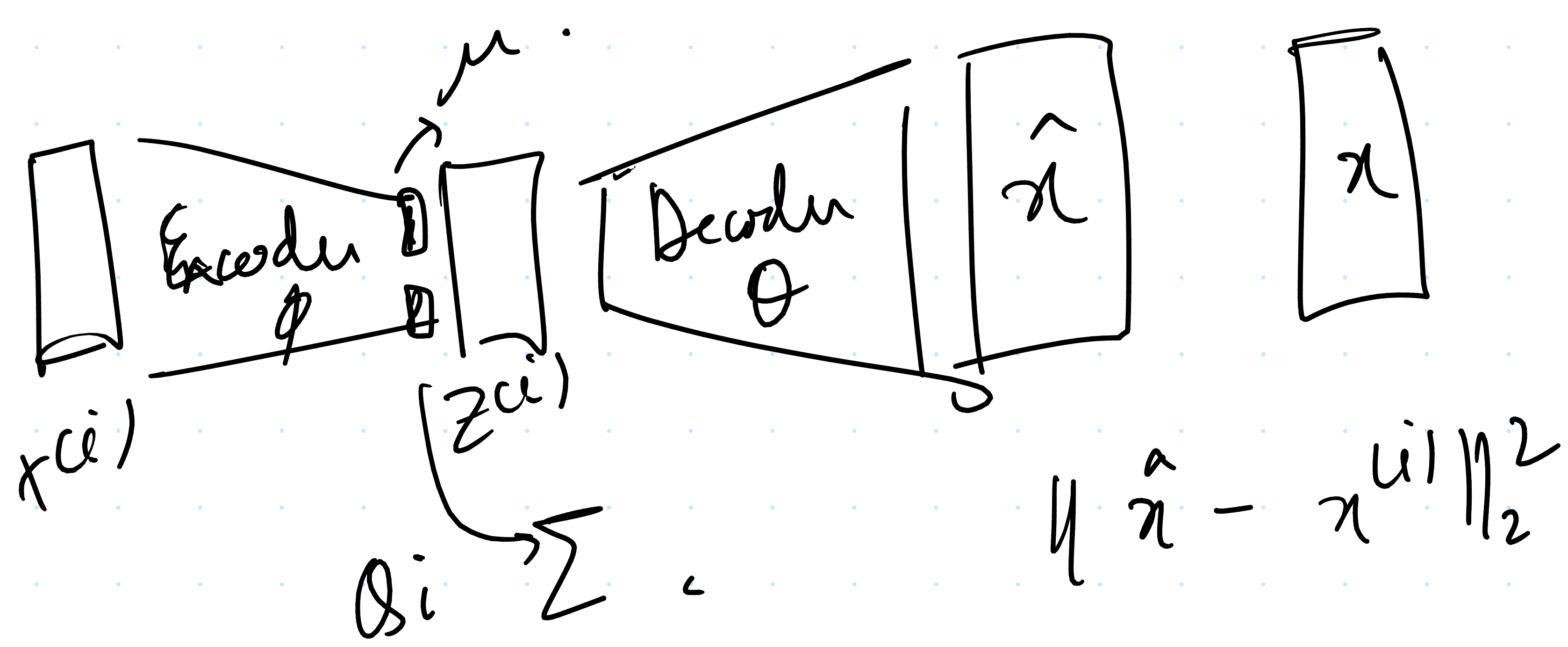
Reparametrization trick.

Maximizing the ELBO,

$ELBO(\phi, \psi, \theta)$

Easy gradient.

Expectation of the Gradients are approximated using Monte-Carlo Estimate.



Maximizing the ELBO - calculating the Gradients and then gradient ascent steps. Alternate to EM $\rightarrow$ fit a latent variable model.