Probability of (z) given (x)

$$p(z|x) = \frac{p(x|z) \cdot p(z)}{p(x)} \quad (1)$$

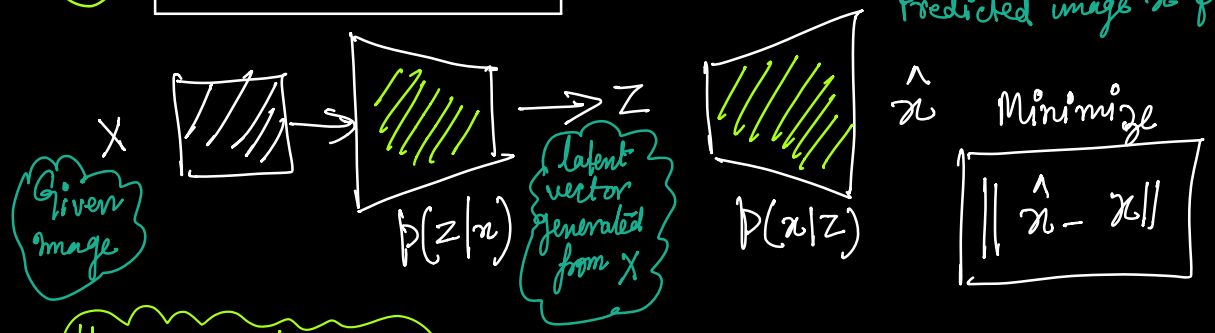
Our Youtube Channel :-

Deep learning for all at Manas lab at IIT Mandi.

Aditya Nigam - I.I.T Mandi.

:) Can be converted into (2) [Bottleneck]

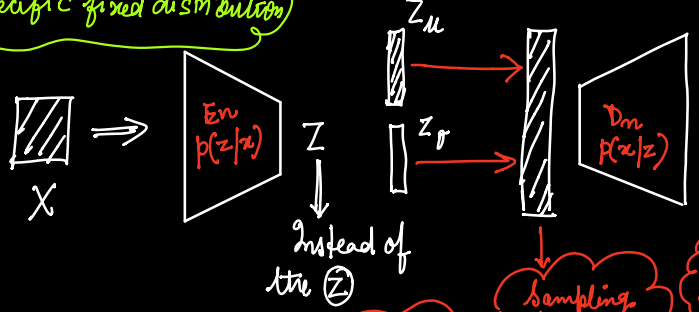
(A) Basic Auto encoder



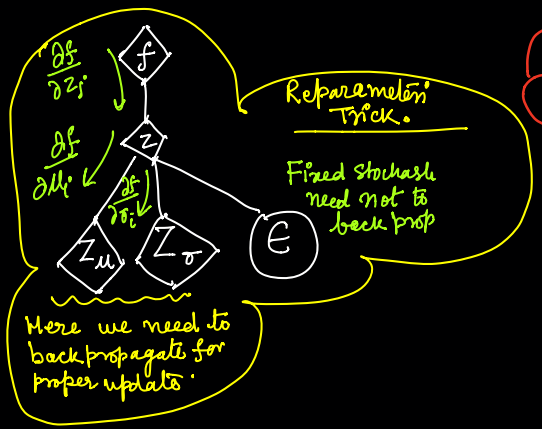
Here our main question is (a) How to make sampling backpropagable using reparameterization technique

(b) Why we want to train encoder so as to produce z's that follow a specific fixed distribution

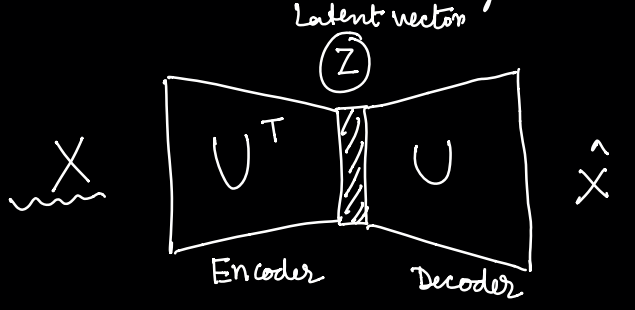
(B) Variational Autoencoders



$$KL\text{-}Div(G(z_\mu, z_\sigma), N(0,1)) = \frac{1}{2}\sum_{i=1}^{n}\left(\mu_i^2 + \sigma_i^2 - \log(1e\text{-}8 + \sigma^2)\right) - 1$$



In basic auto encoder data goes into the bottleneck and reconstructed



Loss = $\min \|x - \hat{x}\|$ (reconstruction error)

* If there is no non-linearity (i.e w/o any activation fn) and there is only one hidden layer then this is very similar to PCA analysis.

This doesn't ensure that Such A.E and PCA both learns the identical basis but may span the similar space

Encoder (E)

① $Z = U^T X$

$(p \times 1) = (p \times d)(d \times 1)$

$X \in \mathbb{R}^d$ (d-dim vector)

$Z \in \mathbb{R}^p$ (p-dim vector)

Encoder is learning some transformation that can convert (X) to (Z)

i/p — latent vector

Decoder (D)

② $\hat{X} = U Z$

$(d \times 1) = (d \times p)(p \times 1)$

applying ① $\hat{X} = U U^T X$

Hence our loss fn has to be $\min \|X - \hat{X}\|$

$\therefore \min \|X - U U^T X\|$

main difference b/w PCA and A.E can be that in PCA

$U U^T = I$ [U is orthonormal by construction]

In A.E [U] may not be learned as orthonormal.

One can train deep auto encoders with non-linearity in order to learn better representation.

Important Concepts

(A) Bayes Theorum:

⊙ Conditional probability ∴ events are (A) & (B)

$\therefore P(A|B) = \frac{P(A \cap B)}{P(B)}$

$P(B|A) = \frac{P(B \cap A)}{P(A)}$

$\Rightarrow P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

$\Rightarrow \boxed{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}}$

Hypothesis — Evidence — Prior

Assuming 1 student in the class of (20) has flu.
Event A: Student is ABC $\Rightarrow [P(A) = 1/20]$

Prior Probability (w/o any knowledge)

Evidence B: 5 Girls & 15 Boys

Now given this new evidence what is the probability that ABC has flu.

$P(A|B) = \frac{1}{5}$ (goes up) if girl $\Rightarrow$ 0 (if student is a Boy)

New evidences are going to influence the Hypothesis

(B) Information (I): How one can estimate the amount of information in a sentence/expression

$\Rightarrow$ (An event) (X)

Here we can have 3 things:

	X (Event)	P(X) (Probability)	$-\log(P(X))$ (Information)
(1)	Virat Scored a century.	↑ (highly probable event)	↓ (less information)
(2)	Kenya wins Cricket world Cup	↓ (rare event)	↑ (high information)
(3)	Tomorrow it rain or don't	1 (Certain event)	0 [no information]

Since $P(x) \in (0,1)$ hence $(-\log)$.

So basically more events carry more information

(C) Average of Information is Entropy (H):

The expected value of information wrt any event (x) averaged over all values (x) can attain is Entropy (H).

$$H = -\sum p(x) \log p(x)$$

- This is the expected value of $[\log p(x)]$ wrt $p(x)$
- Summation over all x's
- Probability of that (x) to happen
- Information Content in any (x)

(D) KL-Divergence (KL-Div):

In order to compute the similarity between two distributions say (p) and (q) KL-Div $(p \| q)$ can be used defined as the KL-Div of (q) distribution wrt (p).

(i) Entropy of (q) – Entropy of (p)

- Entropy of (q): Amount of information in (q) distribution
- Entropy of (p): Amount of information in (p) distribution

$$-\sum q(x) \log q(x) + \sum p(x) \log p(x)$$

- $-\sum q(x) \log q(x)$: This expectation wrt $q(x)$
- $\sum p(x) \log p(x)$: This expectation is wrt $p(x)$

KL-Div is almost this except that the expectation is always computed wrt $p(x)$ as KL-Div is wrt $p(x)$

(ii) $-\sum p(x) \log q(x) + \sum p(x) \log p(x)$

↓ This is the Cross entropy between (p) and (q) distributions.

↓ This is (true) entropy of (p) distribution

Here both expectations are wrt $p(x)$

Hence, KL-Div $(p(x)|q(x))$ can be formally defined as the difference between average information of $q(x)$ wrt $p(x)$ and that of $p(x)$ wrt $p(x)$.

$$\text{KL-Div}(p(x)|q(x)) = -\sum p(x) \log q(x) + \sum p(x) \log p(x)$$

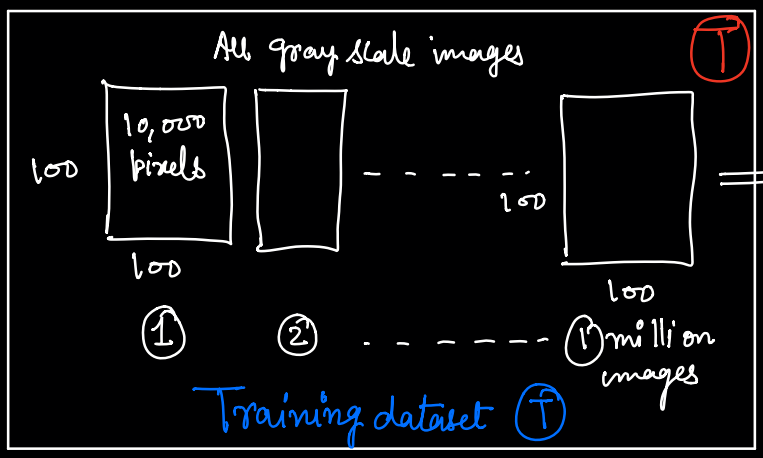
$$= \sum p(x) \log \frac{p(x)}{q(x)}$$

$$= -\sum p(x) \log \frac{q(x)}{p(x)}$$

* KL-Div is not Symmetric as KL-Div $(p|q) \neq$ KL-Div $(q|p)$
  ↳ Hence it is a distance measure b/w a divergence.
* KL-Div ≥ 0 it is always +ve

$$\therefore \text{KL-Div}(q(z) \,||\, p(z|x)) = -\sum q(z) \log \frac{p(z|x)}{q(z)}$$

(We will come back to this.)



Let us assume that we have a very complex and huge training dataset (T) of 1 million images of size $100 \times 100 = 10,000$ pixels per image.

Each image can be seen as 10K pixels each being sampled independently from {0-255}. Hence there can be $(256)^{10,000}$ possible such images. Very big

But any 100×100 image is not just any random 2D matrix and all gray values are not equiprobable at each location. In an image pixel's have dependence, specific gray level co-occurances patterns.

we are assuming this pixel by pixel image sampling (i.e gray value selection) experiment as a stochastic process and can be modeled using random variables.

Collection of random variables where each of them uniquely associated with an element in the set

$$\therefore \quad P(X) = P(X_1, X_2, X_3 \sim \cdots X_{10,000})$$

Joint probability distribution

Depending upon our training dataset (T), we wanted to estimate $P_\theta(x)$ where $P_\theta(x)$ = Probability of $(X \in T)$, with the distribution parameterized over $[\theta]$.

$$\theta^* = \arg\max_\theta \left[P(x \in T) \right]$$

minimizing the rest.

But why we are interested to compute $P_\theta(X)$

- (a) Classification: Helps us to discriminate b/w images that are coming from (T) or rest.
- (b) Generative modeling: It can help us to sample new/unseen X_i^s from $P_\theta(X)$ distribution that are not even present. Such as non-trivial views, poses, interpolation b/w 2 views/poses.

For this image sampling experiment $P(X) = P(X_1, X_2 \cdots X_{10,000})$ is multivariate probability distribution. If we estimate it we know how to sample a new image (basically 10,000 values) from this joint distribution.

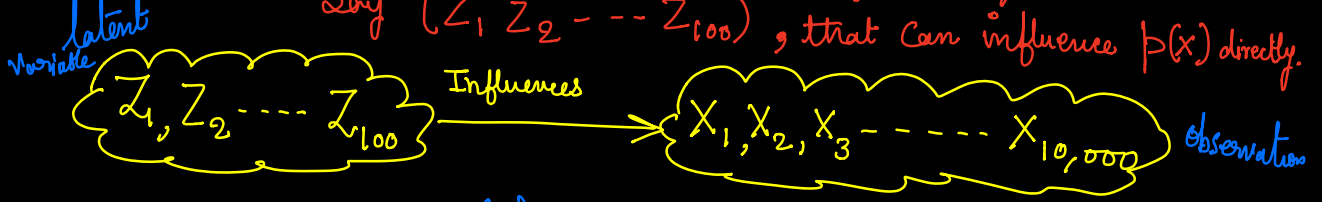
But such probability distribution estimation is intractable and very complex.

$$P(X) = P(X_1, X_2 \cdots X_{10,000}) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1, X_2) \cdots \cdots P(X_n|X_1, X_2 \cdots X_{n-1})$$

(M1)

- Computation is infeasible.
- Since (X) is an image these X_i^s are not independent
- There is huge amount of dependency between random variables (X_i^s)

we can assume another set of hidden/latent variables Say $(Z_1, Z_2 \cdots Z_{100})$, that can influence $P(X)$ directly.



Now our observation (X) got dependent upon latent variables (Z)

Basically our image (X), got influenced by few factors (Z), such as pose, illumination, noise - - - . (8)

Since dimensions of (Z) is far lesser than (X) it is easier to get hold of $P_\theta(x)$ via (Z)

for example our line dataset, length, color, angle, width ...

external Parameters

⊛ We have observed some (X) that is influenced by (Z)

(M2) ∴ Distribution of Observed data (X) is significantly influenced by (Z)

$$p(x) = \int_z p(x,z)\,dz = \int_z p(x|z)\cdot p(z)\,dz$$

(marginalized over all z's)

Since this marginalization has to happen over all z's

(z's can also be huge) may be dimentionally lesser than x

[This is also intractable]

Hence this marginalized estimation is also not feasible

Now both M_1 & M_2 are not feasible, let us try some thing else that is convinient

Bayes Theorem:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

- $P(H|E)$: Hypothesis whose probability may be affected by evidence (E) i.e data — Probability of Hypothesis given the observed evidence — Posterior
- $P(E|H)$: Probability of observing (E) given (H) — Likelihood
- $P(H)$: Prior probability distribution w/o evidence / data
- $P(E)$: Marginal likelihood (Remains to be the same for all possible Hypothesis, (data is fixed))

$x = x_1 \dots x_N$ — observed variable

$z = z_1 \dots z_M$ — set of latent variable

with $N >> M$

and

$p(z, x)$ is joint distribution

It can be seen as a mapping $x \to z$

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

Given any x, we need to estimate a suitable (z) that can be used as a bottle neck of the observed (x)

This makes the real estimation again intractable (exponential in time) w/o any close form

This is again marginalized over our latent variable

$$p(x) = \int_z p(z,x)\, dz$$

Instead of statistically estimating $p(z|x)$ we wanted to approximate $p(z|x)$ via. $q(z) \in \mathcal{L}$ → family of some known tractable distribution (say unit normal, Gaussian ---)

$$\therefore \quad q^*(z) = \arg\min_{q(z) \in \mathcal{L}} \text{KL-Div}(q(z) \,|\, p(z|x))$$

$p(z|x)$ — We wanted to approximate this mapping function

Via.

$q(z)$ — This is a distribution choosen from the family of tractable distribution (easy to estimate)

(*) It is choosen to be tractable as well as close to $p(z|x)$

(8) we need to estimate the best set of parameters of (9) that make it as close as above condition.

$$\therefore \quad \text{KL-Div}(q(z) \,|\, p(z|x)) = -\sum_z q(z) \log \frac{p(z|x)}{q(z)}$$

(*) We wanted to maximize $P_\theta(X)$ that seems intractable. (10)
Instead we started to approximate $p(z|x)$ via $q(z) \in \mathcal{L}$ (family of tractable fns)

[WHY]

$$KL\text{-}Div\big(q(z) \,|\, p(z|x)\big) = -\sum_z q(z) \log \frac{p(z|x)}{q(z)}$$

$$= -\sum_z q(z) \log \frac{p(x,z)}{p(x) \cdot q(z)}$$

Remember estimation has been carried out for a given x hence x is fixed.

$$= -\sum_z q(z) \left[\log \frac{p(x,z)}{q(z)} - \log p(x) \right]$$

$$= -\sum_z q(z) \log \frac{p(x,z)}{q(z)} + \log p(x) \sum_z q(z) \quad \to (1)$$

$$= -\sum_z q(z) \log \frac{p(x,z)}{q(z)} + \log p(x)$$

$$\log p(x) = KL\text{-}Div\big(q(z) \,|\, p(z|x)\big) + \sum_z q(z) \log \frac{p(x,z)}{q(z)}$$

for any given x i.e data, no matter what is z { that we are trying several hypothesis }
$\log p(x) = \text{Const}$

KL-Div is always +ve ≥ 0

$\mathcal{L}$: Variational lower bound of $p(x)$ is the evidence
This will be called as ELBO: Evidence lower Bound

Minimize (B)

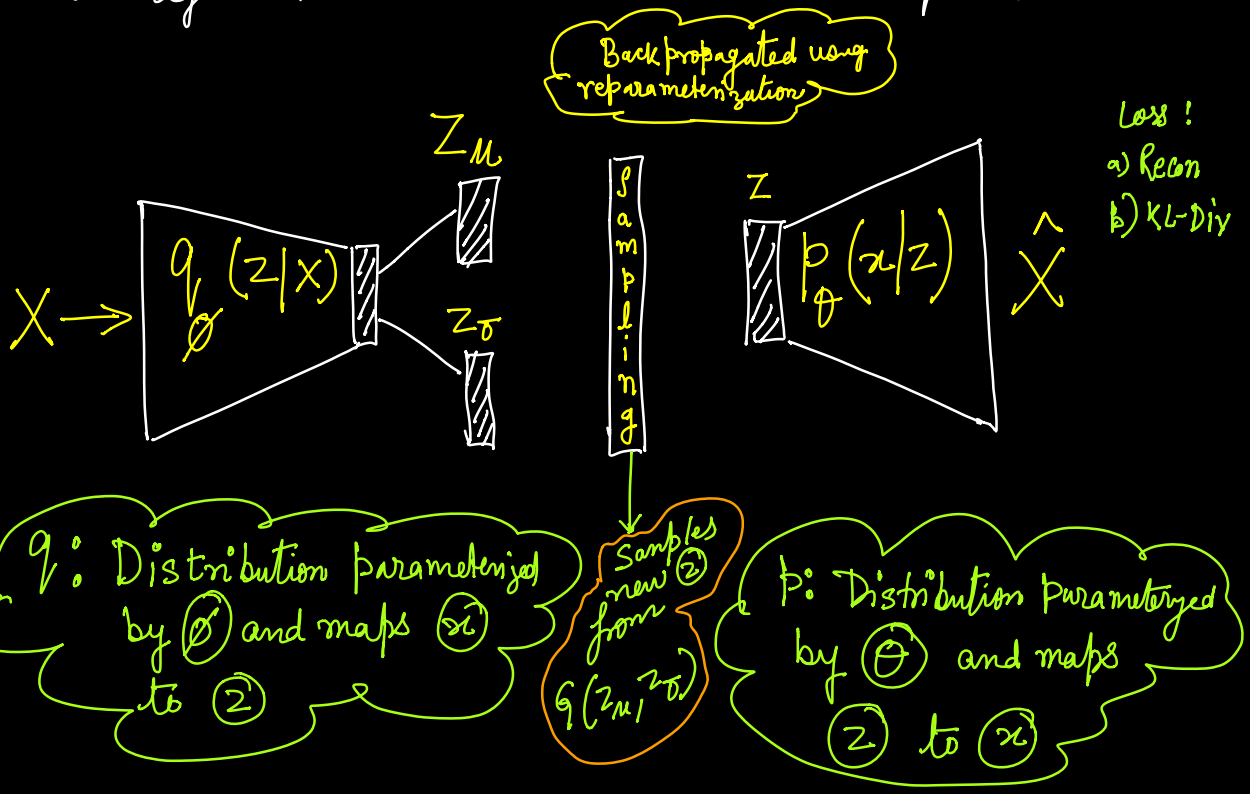
Constant (A)

as a result this Maximized (C)

Maximizing (C) is equivalent to minimizing (B)

This also mean that maximizing the lower bound of a fn implies we are maximizing the fn [$\log p(x)$].

Let us again redraw variational auto encoder diagram :



ELBO

$$L(\theta, \phi; x, z) = \underbrace{E_{q_\phi(z|x)} \log p_\theta(x|z)}_{\text{Reconstruction}} - \underbrace{D_{KL}(q_\phi(z) \| p(z))}_{\text{KL- Div}}$$

This need to maximized the ELBO

Maximized

Minimized

Known tractable distribution

$N(0,1)$

This is the log likelihood of x given z. Maximizing it ensures z should have to generate x.

But these z's are coming from Encoder hence averaged over $q_\phi(z|x)$

Minimizing this KL-Div makes $q_\phi(z)$ as close to $p(z) = N(0,1)$

Just copy ELBO (L) from above:

$$L = \sum_{z} q(z) \log \frac{p(x,z)}{q(z)}$$

Evidence lower bound

$(z \in q_\phi(z|x))$

$$= \sum q(z) \log \frac{p(x|z) * p(z)}{q(z)}$$

$$= \sum q(z) \log p(x|z) + \sum q(z) \log \frac{p(z)}{q(z)}$$

$$= \underset{(A)}{E} \log p_\theta(x|z) - \underset{(B)}{KL\text{-}Div}\left(q(z) \,|\, p(z)\right)$$

$q_\phi(z|x)$ → Encoder

$p_\theta(x|z)$ → Decoder

Log likelihood of x given z averaged of z's sampled from $q_\phi(z|x)$

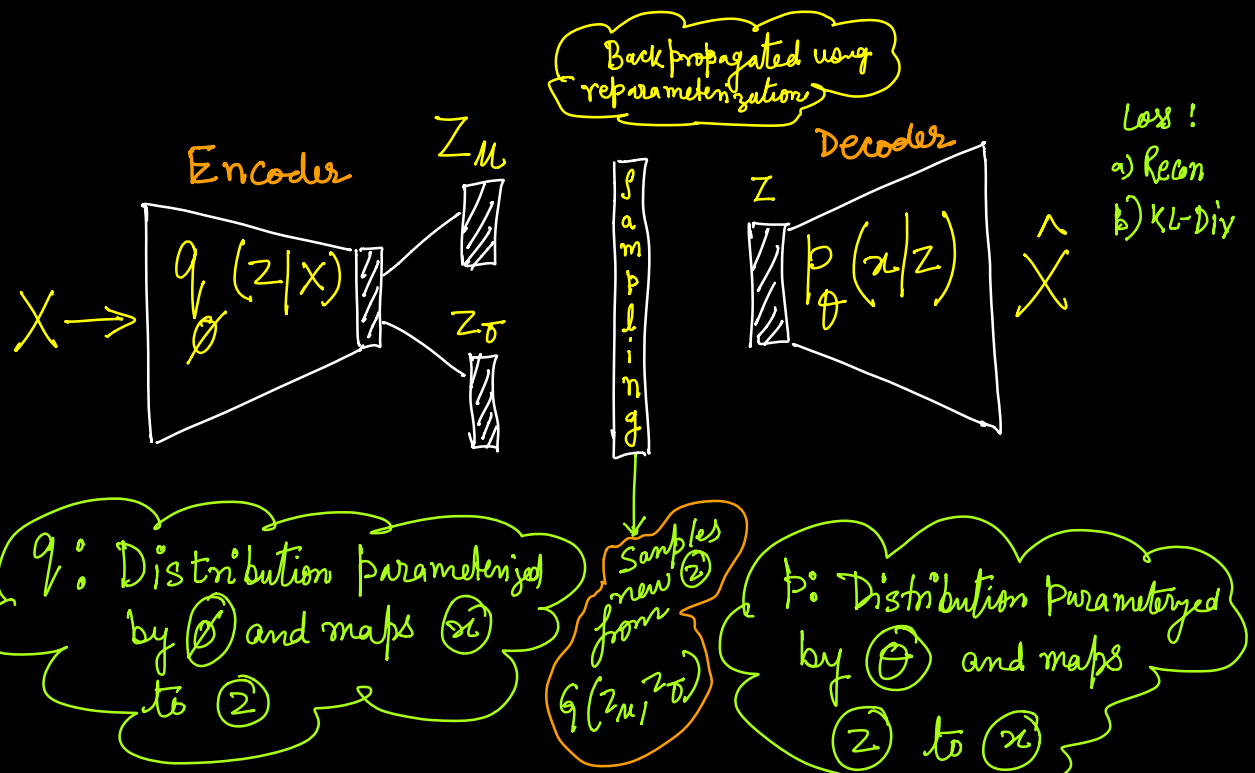
Minimizing this KL-Div ensures that $q(z)$ moves as close as possible to $N(0,1)$.

We fix this distribution to some known and tractable distribution such as $N(0,1)$

Still ELBO has one term to be maximized (A) ↑ and other to be minimized (B) ↓ (ELBO itself have to be Maximized). Hence converting it to a loss fn that can be minimized optimally.

$$\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]$$

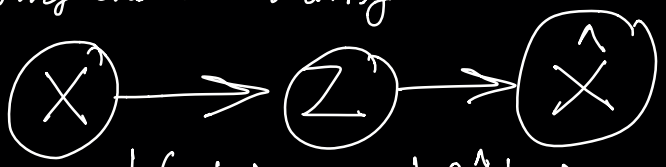
① It is log likelihood of (x) given (z),
② Averaged over (z)'s,
③ Sample from $q_\phi(z|x)$



∴ This **likelihood maximization** can be suitably realized as reconstruction **loss minimization**

① Get a sampled (z) from given (X) (via. Encoder $q_\phi(z|x)$)
② Feed (z) to Decoder (i.e $p_\theta(x|z)$) and get $\hat{X}$.
③ In order to maximize the above averaged log likelihood, we expect that $\hat{X}$ must be as close to (X).

∴ Observing the transitivity.



$\longrightarrow p(z|x) \qquad p(\hat{x}|z)$

$\longrightarrow$ This ultimately becomes $p(\hat{x}|x)$

assuming this to be a gaussian distribution

$$\Rightarrow p(\hat{x}|x) = e^{-\|x-\hat{x}\|}$$

taking log both sides

(A) $\log p(x|z) \longrightarrow \log p(\hat{x}|x) \longrightarrow \log\left(e^{-\|x-\hat{x}\|}\right)$

$$\Downarrow$$

$$-\|x-\hat{x}\| \quad (B)$$

Maximizing (A) is equivalent to minimizing $-(B)$

$$\|x-\hat{x}\|_2$$

Hence final VAE loss:

$$L(\phi, \theta; x, z) = \min \|x-\hat{x}\|_2 + KL\left(q(z)\,|\,p(z)\right)$$

- $\min$: Averaged over z's sampled from the distribution predicted by $q_\phi(z|x)$ [ENCODER]
- $\|x-\hat{x}\|_2$: Fed back to $p_\theta(x|z)$ [DECODER] to get an $\hat{x}$.
- $q(z)$: Distribution of z's sampled from $\mathcal{G}(z_\mu, z_\sigma)$
- $p(z)$: $\mathcal{N}(0,1)$

RECAP :

(1) First we observed that estimating $p(x)$ directly is Intractable.

$$p(x) = p(x_1, x_2 \text{——} x_{10,000}) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2) \cdot \dots \cdot p(x_n|x_1, x_2 \dots x_{n-1})$$

(2) Then we try to estimate $p(x)$ indirectly via latent variables (z). But that too observed to be Intractable.

$$p(x) = \int_z p(x, z)\, dz \quad \text{(marginalized over all } z\text{'s)} \quad = \int_z p(x|z) \cdot p(z)\, dz$$

Since this marginalization has to happen over all z's (z's can also be huge / may be dimentionally lesser than x) [This is also intractable]

Hence this marginalized estimation is also not feasible

(3) Later we tried to estimate $p(x)$ via. $p(z|x) = \frac{p(z,x)}{p(x)}$ which was also not tractable due to $p(x)$ marginalization.

(4) Then we decided to approximate $p(z|x)$ via. $q(z) \in \mathcal{L}$ {family of tractable fn}

$$q^*(z) = \underset{q(z) \in \mathcal{L}}{\arg\min}\ KL\text{-}Div\left(q(z) \,|\, p(z|x)\right)$$

$p(z|x)$: we wanted to approximate this mapping function

Via.

$q(z)$: This is a distribution choosen from the family of tractable distributions (easy to estimate)

(*) It is choosen to be tractable as well as close to $p(z|x)$

(*) we need to estimate the best set of parameters of (q) that make it as close as above condition.

(5) The minimization of above KL-Div has been observed to be equivalent to maximizing ELBO($\mathcal{L}$)

$$\log p(x) = KL\text{-}Div\left(q(z) \,|\, p(z|x)\right) + \sum_z q(z) \log \frac{p(x,z)}{q(z)}$$

($\mathcal{L}$): Variational lower bound of $p(x)$ is the evidence. This will be called as ELBO: Evidence lower Bound

for any given (x) i.e data, no matter what is (z) {that we are trying several hypothesis}, $\log p(x)$ = Constant

Constant (A)

KL-Div is always +ve ≥ 0

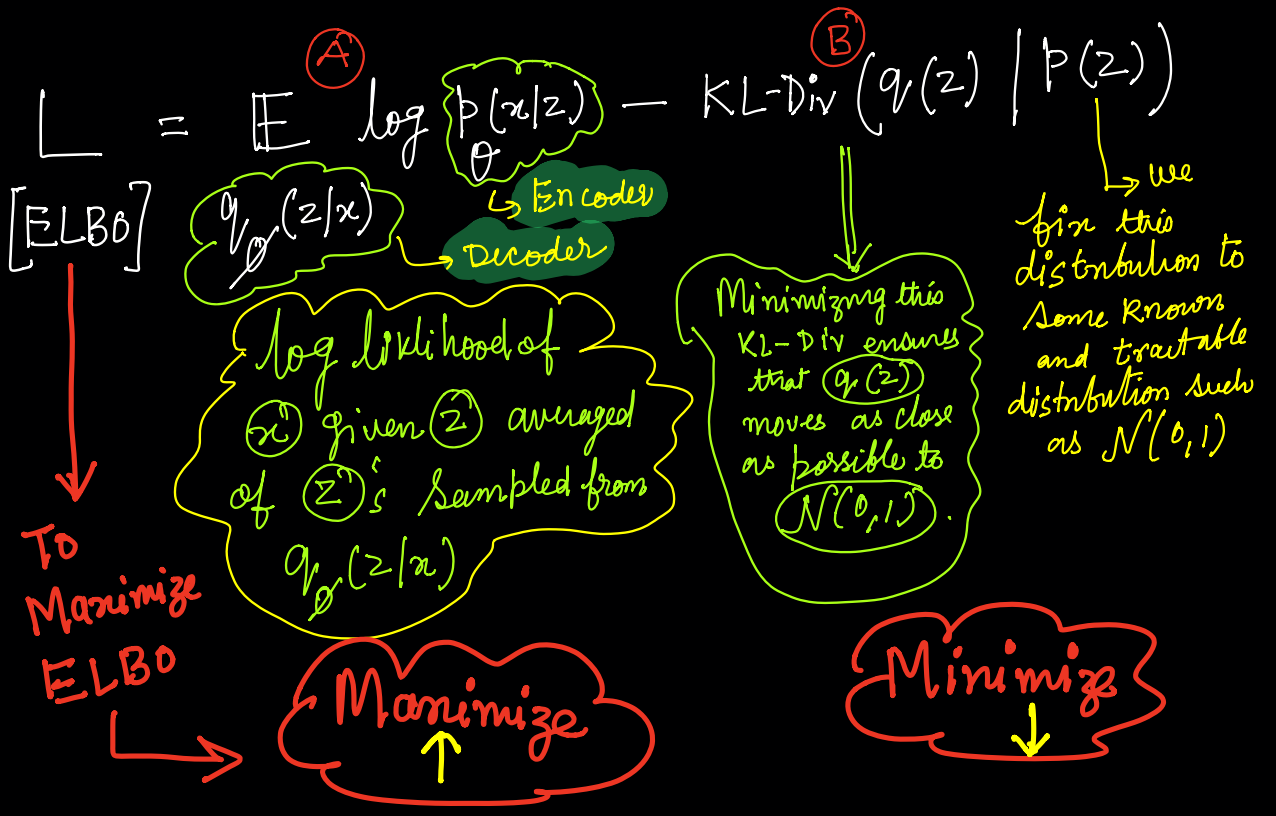
Minimize (B)

as a result this Maximized (C)

Maximizing (C) is Equivalent to minimizing (B)

This also mean that maximizing the lower bound of a fn implies we are maximizing the fn [$\log p(x)$].

(6) ELBO maximization is equivalent to KL-Div($q_\phi(z|x)$ | $p(z)$) minimization and $\mathbb{E}\left[\log p_\theta(x|z)\right]$ maximization $q_\phi(z|x)$



(7) $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ maximization is equivalent to minimize $\|X - \hat{X}\|_2$.

(A) $\log p(x|z) \longrightarrow \log p(\hat{x}|x) \longrightarrow \log\left(e^{-\|x-\hat{x}\|_2}\right)$

$\Downarrow$

$-\|X - \hat{X}\|_2$ (B)

Maximizing (A) is equivalent to minimizing (−B)

$\|X - \hat{X}\|_2$

(8) Finally the loss function of VAE is

$$L(\phi, \theta; X, z) = \min \|X - \hat{X}\|_2 + KL\left(q(z) \,\middle|\, p(z)\right)$$

- $\min$: Averaged over z's sampled from the distribution predicted by $q_\phi(z|x)$ [ENCODER]
- $\|X - \hat{X}\|_2$: Fed back to $p_\theta(x|z)$ [DECODER] to get an $\hat{X}$.
- $q(z)$: Distribution of z's sampled from $q(z_\mu, z_\sigma)$
- $p(z)$: $N(0,1)$

Final Story :

(A) Let us assume that our observation is MNIST etc.

(B) It is a huge and varying dataset hence the images (X) must be following a very complex distribution (P(X))

(C) Using Encoder we try to reduce the dimentionality of (X) (to (Z) latent variables), so that reduced dimentional latent vectors (Z) follow some known tractable distribution.

(D) Subsequently we also train a Decoder network that can sample latent vectors (Z) from Gaussian distributions and generate images similar to MNIST.

(E) Finally after training Encoder is discarded. we just need to sample (Z)'s from $N(0,1)$ and feed it to Decoder that will generate noval / new images similar to MNIST but not already present in it.