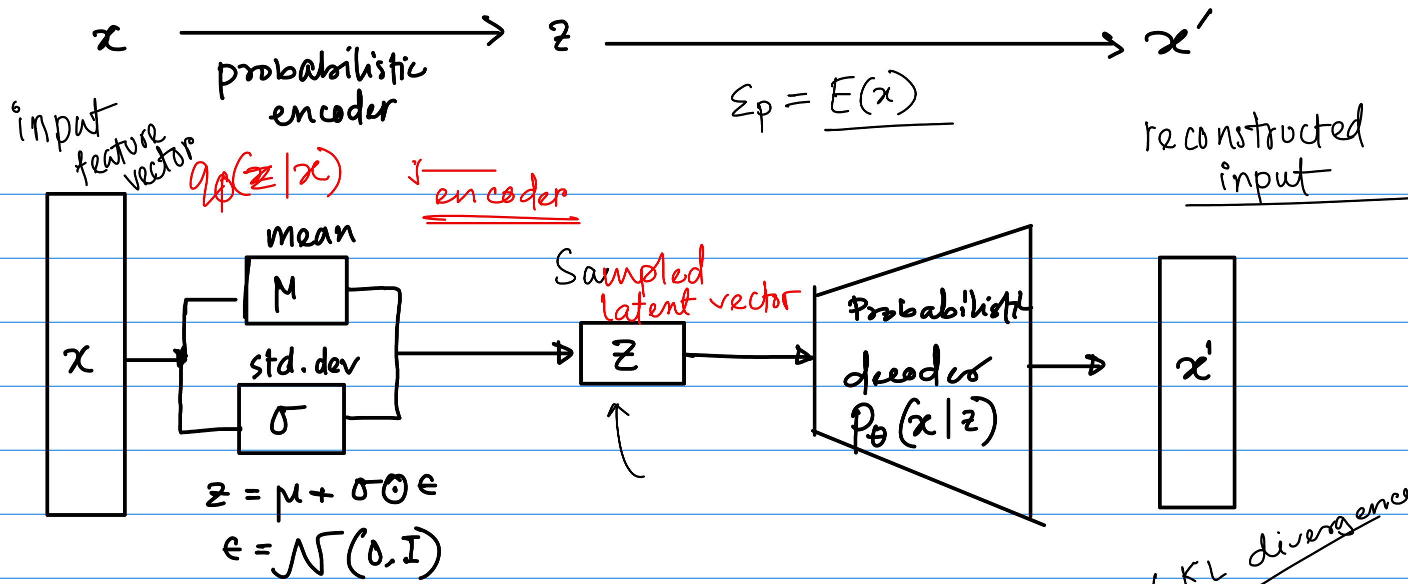
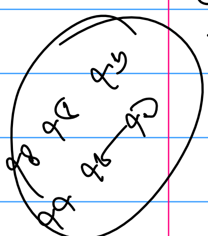


KL divergence

Loss function : $L(\theta, \phi) = -E_{z \sim q_\phi(z|x)}\left[p_\theta(x|z)\right] + D_{KL}\left(q_\phi(z|x) \| p_\theta(z)\right)$

q_ϕ $\quad$ p_θ

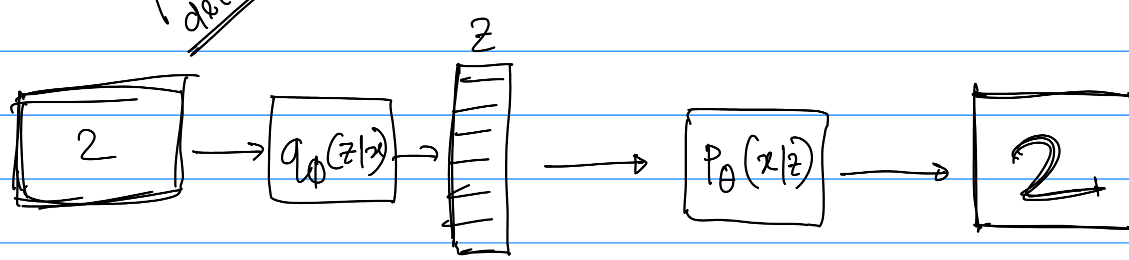
Goal of Variational Autoencoder :

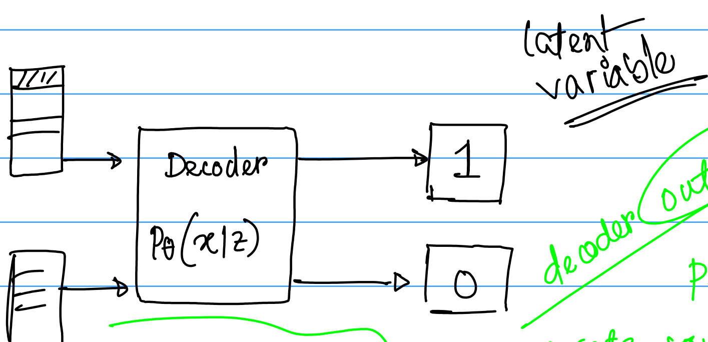


To find $q_\phi(z|x)$ of some latent variable which we can sample/ or generate from $\boxed{q_\phi(z|x)}$ in order to generate new samples

$x' \sim \boxed{p_\theta(x|z)}$ $\quad$ not present in the input

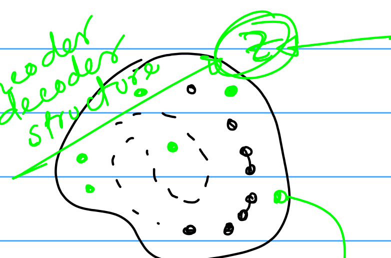
decoder





decoder output

encoder decoder structure



$p_\theta(z)$

generate some new image

But ~~VAE~~ don't know

z

~~$p_\theta(z|x)$~~

$q_\phi(z|x)$

vae dont know if the latent space vector z is valid or not.

it forces the encoder to produce a probability distribution function and then try to make them further

Probability Concepts Required:

$p(x)$: prob of a random var x

$p(x|y)$: prob of x given y has happened.

$$p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)}$$

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

$p(y) \rightarrow$ prior

$\frac{p(x|y)}{p(x)} \rightarrow$ likelihood ratio

posterior probability $= \frac{P(x,y)}{P(x)}$ joint prob dist

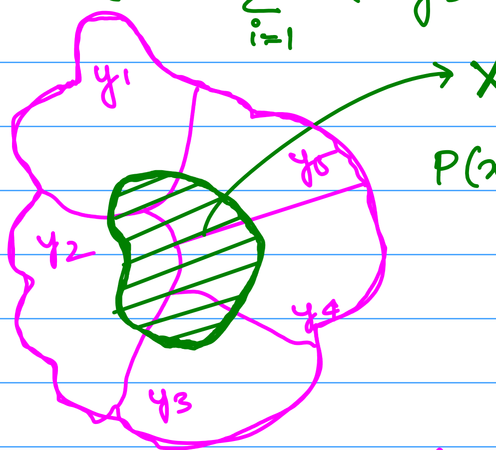
∴ Total probability theorem : $y_1 \; y_2 \; y_3 \; \cdots \; y_n$ (no overlap)

mutually exclusive $(y_i \cap y_j = 0)$

Event X, union of multiple mutually exclusive events :

$$P(X) = \sum_{i=1}^{N} P(x|y_i)\, P(y_i) \longrightarrow (2)$$

$$P(y|x) = \frac{P(x,y)}{P(x)} \longrightarrow (1)$$



$$P(x) = \sum_{i=1}^{5} P(x, y_i)$$

$$= \sum_{i=1}^{5} P(x|y_i)\, P(y_i) \longrightarrow (2)$$

$$p(y|x) = \frac{P(x|y)\, P(y)}{\boxed{\sum_{i=1}^{N} P(x|y)\, P(y_i)}}$$

$\boxed{\sum_{i=1}^{N} P(x|y)\, P(y_i)} \rightarrow P(X)$

$$P(y|x) = \frac{P(x|y)\, P(y)}{P(x)}$$

$$= \frac{P(x|y)\, P(y)}{\sum_{i=1}^{N} P(x|y_i)\, P(y_i)}$$

Expectation of a Random Variable:

$$E(x) = \sum_{i=1}^{K} x_i \cdot P(x = x_i) = E(x)$$

③
$P(3) = \frac{1}{6}$
$P(3) = 1/6$

$p(x=3 \mid y \text{ is odd})$

$$P(x \mid y) = \frac{P(y \mid x)\, P(y)}{P(x)}$$

$=$

For fair die: $\{1,2,3,4,5,6\}$
$P = 1/6$

$E(x) =$

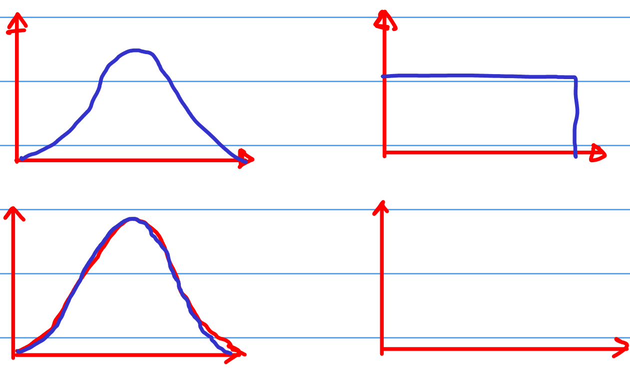
$$E(x) = \sum_{i=1}^{6} x_i \cdot P(x_i)$$

$$= \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6}$$

$$= \frac{21}{6} =$$

$$= \frac{21}{6} =$$

KL-Divergence: Kullback Divergence: 2020



high value

low value

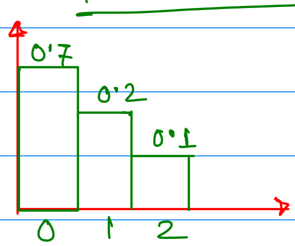
Measure of how one probability is different from the other: for two probability distribution P and Q

$$D_{KL} = (P \| Q) = \sum_{x} P(x_i = x) \log \frac{P(X = x)}{Q(X = x)}$$

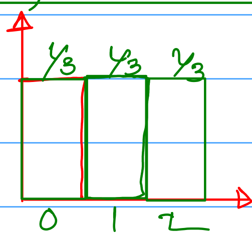
$$= \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \qquad D_{KL}$$

Example:

P distribution



Q distribution:



uniform distribution

Here $D_{KL}(Q \| P) = \sum_{x=0}^{2} Q(x) \log \frac{Q(x)}{P(x)}$

$$= \frac{1}{3} \log\left(\frac{1/3}{0.7}\right) + \frac{1}{3} \log\left(\frac{1/3}{0.2}\right) + \frac{1}{3} \log\left(\frac{1/3}{0.1}\right)$$

$= 0.09673$ nats.

Properties: ① $KL(P \| Q)$ or $KL(Q \| P) \geq 0$
② $KL(P \| Q) \neq KL(Q \| P)$ Non symmetric

\# Consider $P(x) = \mathcal{N}(x; \mu_1, \Sigma_1)$ $\quad$ Σ_1 and Σ_2

$Q(x) = \mathcal{N}(x; \mu_2, \Sigma_2)$ $\quad$ covariance matrix

$$\mathcal{N}(x, \mu, \sigma) = \frac{1}{\sqrt{(2\pi)^K |\Sigma|}} \exp\left(\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Equation of normal distribution:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

then

$$D_{KL} = (P(x) \| Q(x)) = \frac{1}{2}\left[\log \frac{|\Sigma_2|}{|\Sigma_1|} - d + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)\Sigma_2^{-1}(\mu_2 - \mu_1)\right]$$

$\boxed{\text{Proof:}}$

$$KL(P(x) \| Q(x)) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

$$= \sum_x P(x) \cdot [\log P(x) - \log Q(x)] \quad \text{— (1)}$$

and we put the equation of the distribution

$$P(x) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_1|}} \cdot \exp\left(\frac{-(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)}{2}\right)$$

$$\Rightarrow \log P(x) = -\log\sqrt{(2\pi)^K \cdot |\Sigma_1|} - \frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)$$

$$= -\log(2\pi)^{K/2} - \log(|\Sigma_1|)^{1/2} - \frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)$$

$$\log P(x) = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_1|) - \frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) \quad \text{— (2)}$$

Similarly $\log Q(x) = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_2|) - \frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2)$ — (3)

Putting (2) and (3) in (1) we get

$$KL(P(x) \| Q(x)) = \sum P(x)\left[\log(P(x)) - \log(Q(x))\right]$$

$$\sum P(x)\Big[-\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{K}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_2| + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\Big]$$

$$= \frac{1}{2}\sum_x P(x)\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right] \quad \text{— (4)}$$

Considering partially $\sum P(x) = E(x)$ Expected value of x in P

$$\sum P(x) \tfrac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) = E_p\left[\tfrac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right]$$

$$\sum P(x) \tfrac{1}{2} (x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) = E_p\left[\left(\tfrac{1}{2} (x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) \right] \right.$$

Trace : for a square matrix ; trace is the sum of the diagonal elements.

From the properties of trace we know

$$E(X^T A X) = E(tr(X^T A X))$$
$$= E(tr(A X X^T))$$
$$= tr(E(A X X^T))$$

Therefore following this we can rewrite this like the following

$$\tfrac{1}{2} E_p \left[(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) \right]$$

$$= E_p\left[tr\left(\tfrac{1}{2} (x-\mu_1)(x-\mu_1)^T \Sigma_1^{-1} \right) \right]$$

$$= E_p \left[\underbrace{\left[tr\left(\tfrac{1}{2} (x-\mu_1)(x-\mu_1)^T \Sigma_1^{-1} \right) \right]}_{\text{trace of a vector/matrix}} \right]$$

$$= tr\left[\underbrace{E_p\left[(x-\mu_1)(x-\mu_1)^T \right]}_{\text{Covariance matrix}} \times \tfrac{1}{2} \times \Sigma_1^{-1} \right]$$

$$= tr\left[\Sigma_1 \ \tfrac{1}{2} \ \Sigma_1^{-1} \right]$$

$$= tr\left[\tfrac{1}{2} \Sigma_1 \Sigma_1^{-1} \right]$$

$$= tr\left[\tfrac{1}{2} I_k \right]$$

$$= K/2$$

Trace and Expectation trick :

ⓐ ⇒ $E(x) = E(tr(x))$
Since trace of x is scalar

ⓑ ⇒ $tr(AB) = tr(BA)$

ⓑ ⇒ $tr(ABC) = tr(BCA)$
$= tr(CAB)$

But it follows in cyclic order

ⓒ ⇒ $tr(ABC) \neq tr(ACB)$.

ⓓ ⇒ $E(tr(x)) = tr(E(x))$

Considering the second part from equation (04)

$$\sum_x P(x) \left[\tfrac{1}{2} (x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) \right]$$

$$= \sum_x P(x) \left[\tfrac{1}{2} (x-\mu_1) + (\mu_1-\mu_2) \right]^T \Sigma_2^{-1} \left[(x-\mu_1) + (\mu_1-\mu_2) \right]$$

$$= \sum_x P(x) \left[\tfrac{1}{2} (x-\mu_1)^T \Sigma_2^{T} (x-\mu_1) + \tfrac{2}{2} (x-\mu_1)^T \Sigma_2^{-1} (\mu_1-\mu_2) + \tfrac{1}{2} (\mu_1-\mu_2) \Sigma_2^{-1} (\mu_1-\mu_2) \right]$$

$$= E_p \left[\tfrac{1}{2} (x-\mu_1)^T \Sigma_2^{T} (x-\mu_1) + (x-\mu_1)^T \Sigma_2^{-1} (\mu_1-\mu_2) + \tfrac{1}{2} (\mu_1-\mu_2) \Sigma_2^{-1} (\mu_1-\mu_2) \right]$$

Expanding the equation we get the following

$$= \underbrace{E_p\left[\tfrac{1}{2}(x-\mu_1)^T \Sigma_2^{-1}(x-\mu_1)\right]}_{tr\{\frac{1}{2}\cdot\Sigma_2^{-1}\Sigma_1\}} + E_p\left[(x-\mu_1)^T \Sigma_2^{-1}(\mu_1-\mu_2)\right] \rightarrow \underline{\underline{0}}$$

zero

$$+ \underbrace{E_p\left[(\mu_1-\mu_2)^T \Sigma_2^{-1}(\mu_1-\mu_2)\right]}$$

this part is constant due to the absence of x.

$$E_p\left[(x-\mu_1)^T \Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

$$= \left[(E_p(x)-\mu_1)^T \Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

$$= \left[\underbrace{(\mu_1-\mu_1)^T}_{0} \Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

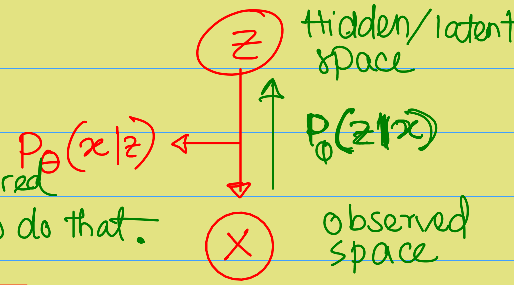
Therefore summing everything up we get

$$KL\left(P(x) \,\|\, Q(x)\right) = \tfrac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - K + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T \Sigma_2^{-1}(\mu_1-\mu_2)\right]$$

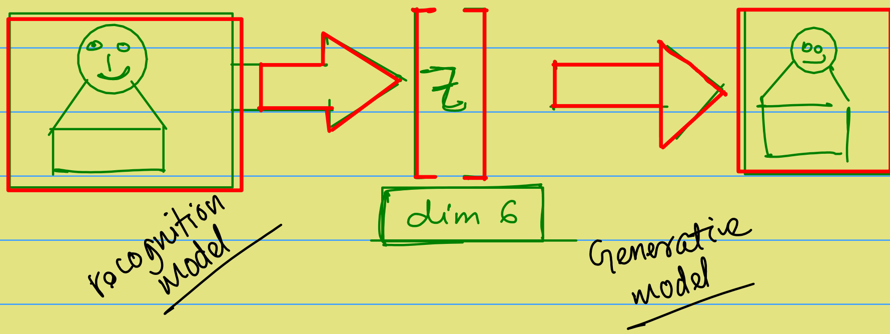
(Proved)

LATENT VARIABLE

$x \longrightarrow z \longrightarrow x'$



latent variable: some feature that are not measured yet. or technolog is not here to do that.

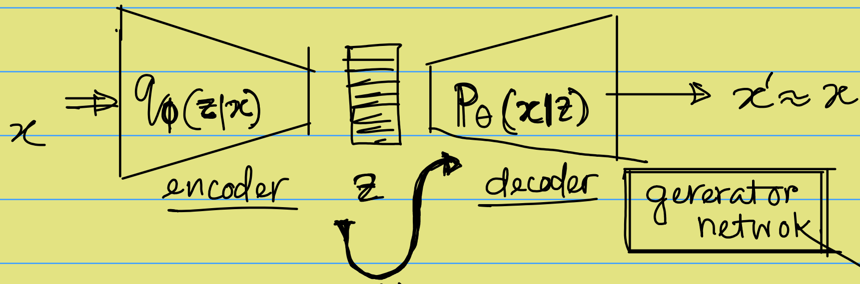


AE : Fixed values

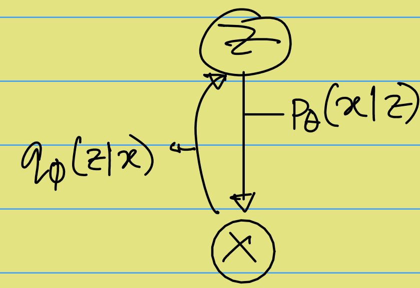
VAE : gives out <u>probability</u>

DERIVATION OF LOSS FUNCTION

to find the distribution of $\boxed{q_\phi(z|x)}$ of some latent variables which can be sampled from $z \sim q_\phi(z|x)$, to generate the new samples x' from $\boxed{P_\theta(x|z)}$



$z^{(1)}$: learned distribution



ϕ θ are learnable parameter.

The problem of approximate inference:

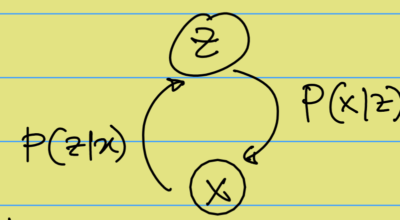
X : set of observed var

z : set of latent variables with joint prob distribution $p(z|x)$.

Aim is to learn the conditional distribution of z given the input set X. ie $P(z|x)$ need to learn :

$$P(z|x) = \frac{P(x|z)\,P(z)}{P(x)} \quad \text{— (A) — equation.}$$

evaluating $P(z|x)$ is difficult because $p(x)$ can not be calculated/ solved.



Reason: $p(x) = \int_z p(x|z)\,p(z)\,dz = \int_z p(x|z)\,dz$ Not available in the closed loop form.

Alternative:

approximate $p(z|x)$ by another dist $q(z|x)$ which is defined in suitable way, done by variational inference. (vi) :

Pose inference problem into optimization problem.

By modeling $p(z|x)$ using $q(z|x)$ where $q(z|x)$ has a simple distribution like gaussian.

calculate D_{KL} between $P(z|x)$ and $Q(z|x)$

$$D_{KL}\left(Q(z|x) \,\|\, P(z|x)\right) = \sum_z Q(z|x) \log \frac{(Q(z|x))}{(P(z|x))}$$

$$= E_{z\sim Q(z|x)}\left[\log \frac{(Q(z|x))}{(P(z|x))}\right]$$

$$= E_{z\sim Q(z|x)}\left[\log(Q(z|x)) - \log(P(z|x))\right] \quad \text{(B)}$$

mentor

By substituting equation (A) in (B) we get

$$D_{KL}\left(Q_\phi(z|x) \,\|\, P_\theta(z|x)\right) = E_z\left[\log(Q_\phi(z|x)) - \log \frac{P_\theta(x|z) \cdot P_\theta(z)}{P_\theta(x)}\right]$$

where $z = z \sim Q_\phi(z|x)$

$$= E_z\left[\log(Q_\phi(z|x)) - \log(P_\theta(x|z)) - \log(P_\theta(z)) + \log(P_\theta(x))\right]$$

Here since expectation is over z, we can bring out $p(x)$ that does not involve z.

$$\therefore D_{KL}\left(Q(z|x) \,\|\, P(z|x)\right) - \log(P(x)) = E_z\left[\log Q(z|x) - \log P(x|z) - \log P(z)\right]$$

Rearranging the equation again, we get

$$\log P(x) - D_{KL}\left[Q(z|x) \,|\, P(z|x)\right]$$

$$= E_z\left[\log P(x|z)\right] - E_z\left[\log Q(z|x) - \log P(z)\right]$$

মুখে নাই কোন কথা

মুখে নাই কোন কথা

$$= \boxed{E_z\left[\log P(x|z)\right] - D_{KL}\left[Q(z|x) \,\|\, P(z)\right]}$$

reconstruction likelihood

VAE objective Function :

And the loss function here is simply loss $= -$ objective function.

$$L(\phi, \theta) = -E_{z\sim q_\phi(z|x)}\left[\log(P_\theta(x|z))\right] + D_{KL}\left[Q(z|x) \,\|\, P(z)\right]$$

Therefore the target of the whole process is to find out the optimal θ and ϕ such that

$$\boxed{\theta^*, \phi^* = \arg\min_{\theta, \phi} L(\theta, \phi)}$$