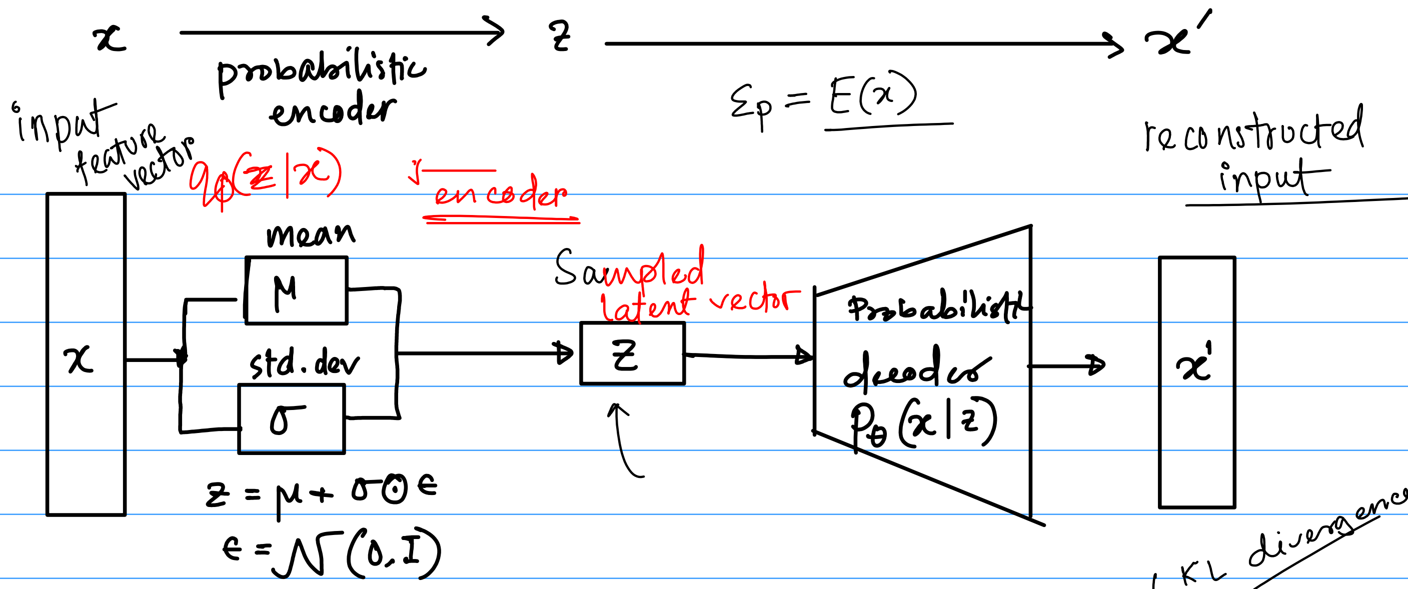
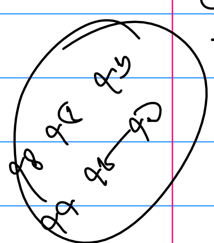


KL divergence

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

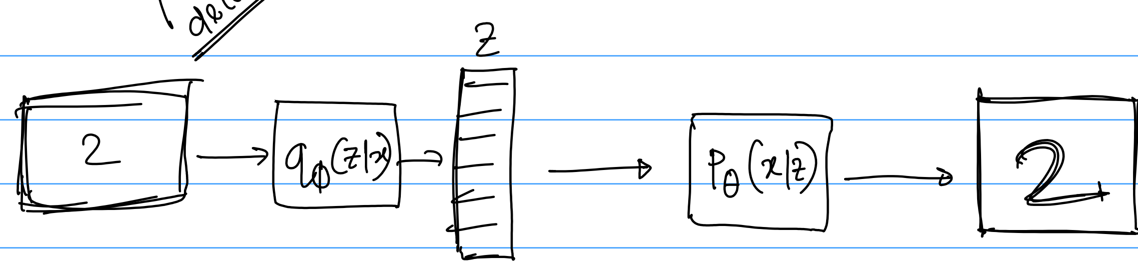
Goal of Variational Autoencoder :

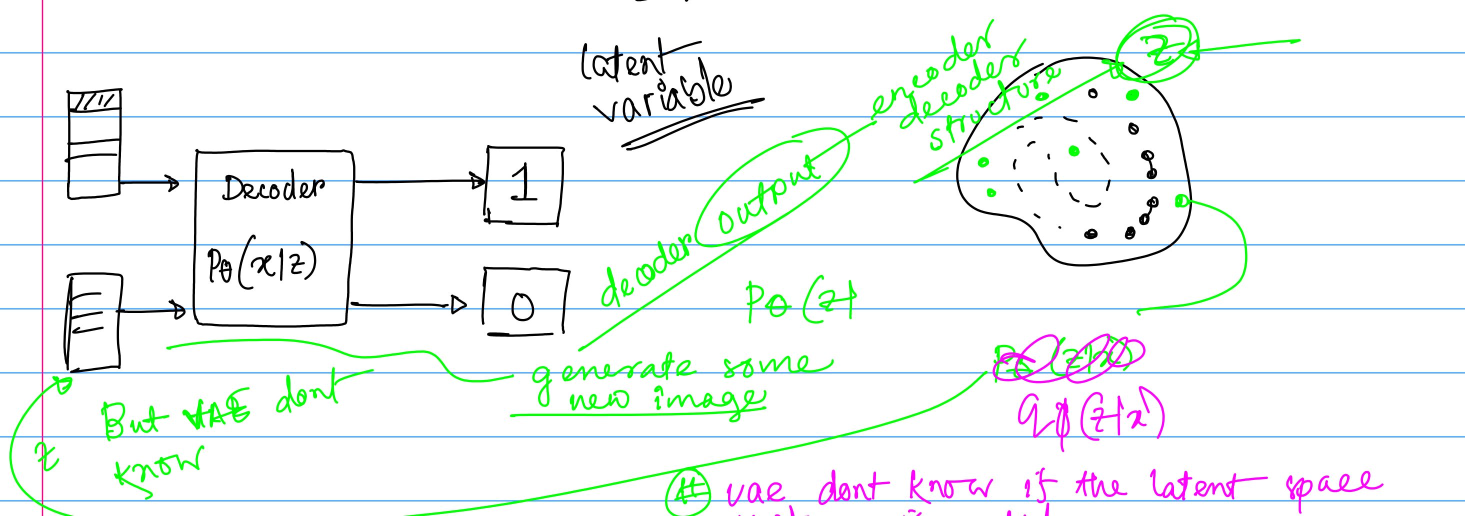


To find $q_\phi(z|x)$ of some latent variable which we can sample/ or generate from $\boxed{q_\phi(z|x)}$ in order to generate new samples (not present in the input)

$x' \sim \boxed{p_\theta(x|z)}$

decoder





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.

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

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

$P(y)$ → prior

$\frac{P(x|y)}{P(x)}$ → likelihood ratio

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

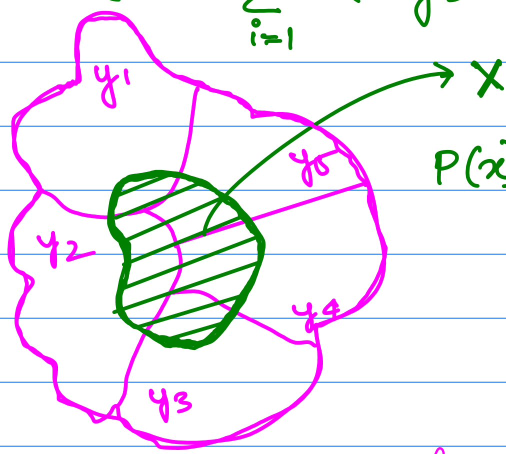
∴ Total probability theorem : $y_1 \; y_2 \; y_3 \; \ldots \; y_n$ (no overlap)

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

Event X, union of multiple mutually exclusive events :

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

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



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

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

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

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

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

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

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

$E(x) =$

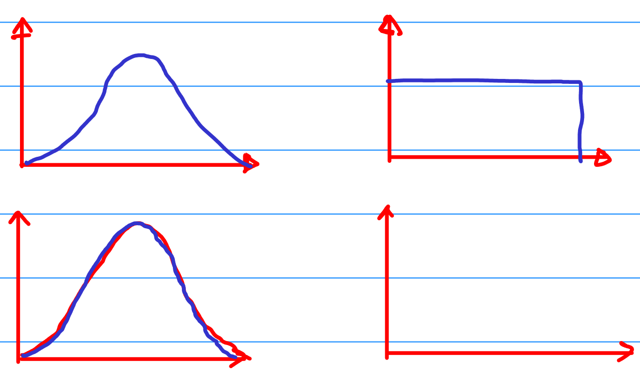
$$E(x) = \sum_{i=1}^{6} x_i \cdot P(x_i)$$

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

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

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

KL-Divergence: Kullback Divergence: 2020



high value

low value

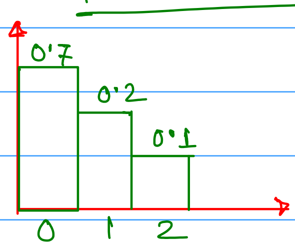
Measure of how one probability is different from the other: for two probability distribution P and Q

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

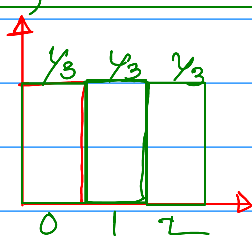
$$= \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \quad D_{KL}$$

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 symetric

Consider $P(x) = \mathcal{N}(x; \mu_1, \Sigma_1)$
$Q(x) = \mathcal{N}(x; \mu_2, \Sigma_2)$

Σ_1 and Σ_2 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

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

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

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

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

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

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

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

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

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

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

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

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

$$E(X^TAX) = E(tr(X^TAX)) = E(tr(AXX^T)) = tr(E(AXX^T))$$

Therefore following this we can rewrite this like the following

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

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

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

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

Trace and Expectation trick:

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

ⓑ $\Rightarrow tr(AB) = tr(BA)$

ⓑ $\Rightarrow tr(ABC) = tr(BCA) = tr(CAB)$

But it follows in cyclic order

ⓒ $\Rightarrow tr(ABC) \neq tr(ACB)$.

ⓓ $\Rightarrow E(tr(x)) = tr(E(x))$

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

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

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

$$= K/2$$

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

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

$$+ \underbrace{E_p\left[(\mu_1-\mu_2)^T \Sigma_2^{-1} (\mu_1-\mu_2)\right]}_{\text{this part is constant}}$$

due to the absence of x.

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

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

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

Therefore summing everything up we get

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

(Proved)

LATENT VARIABLE

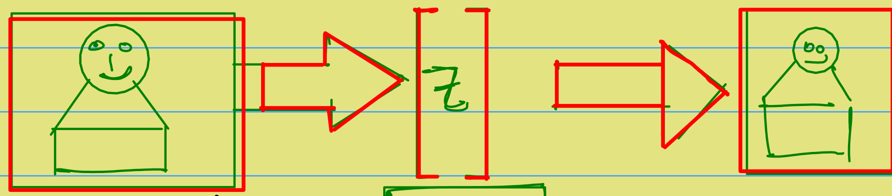
$x \longrightarrow z \longrightarrow x'$

z Hidden/latent space

$P_\theta(x|z)$ $P_\phi(z|x)$

x observed space

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



dim 6

recognition model

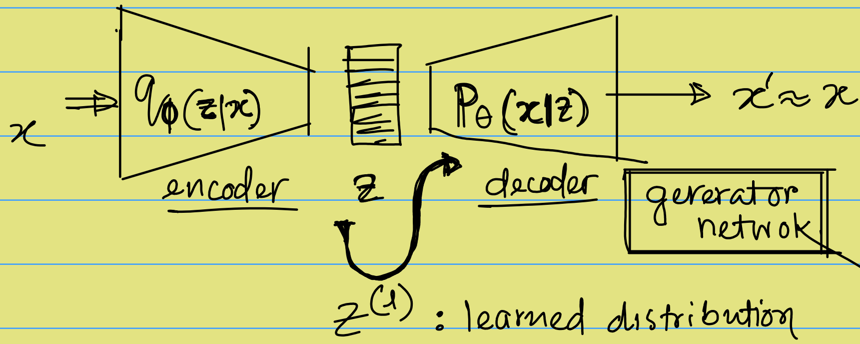
Generative model

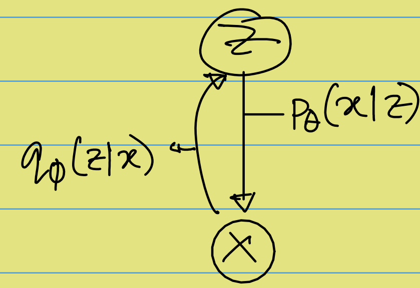
AE : Fixed values

VAE : gives out probability

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 $\boxed{p_\theta(x|z)}$





ϕ θ are learnable parameter.

The problem of approximate inference:

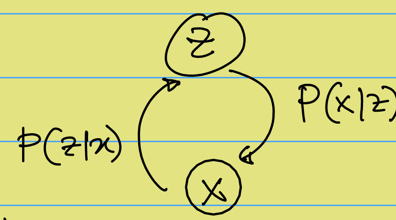
X : set of observed var

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.

$\boxed{\text{Alternative:}}$

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.

z is being sampled from $Q_{\phi}(z|x)$

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

Rearranging the equation again, we get

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

$$= E_{z}\left[\log P_{\theta}(x|z)\right] - E_{z}\left[\log Q_{\phi}(z|x) - \log P_{\theta}(z)\right]$$

মুখে নাই কোন কথা

মুখে নাই কোন কথা

$$= \boxed{\underbrace{E_{z}\left[\log P_{\theta}(x|z)\right]}_{\text{reconstruction likelihood}} - D_{KL}\left[\overbrace{Q_{\phi}(z|x)}^{\text{learned distribution}} \,||\, \underbrace{P_{\theta}(z)}_{\text{prior distribution}}\right]}$$

VAE objective Function :

And the loss function here is simply loss = − objective function.

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

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

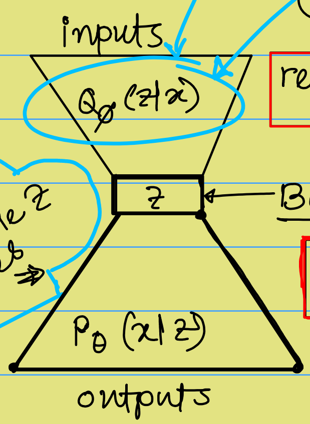
$$\boxed{\theta^*, \phi^* = \underset{\theta, \phi}{\arg\min}\ L(\theta, \phi)}$$

auto regularizer

$$\underbrace{L(\theta,\phi)}_{\text{network parameters to learn}} = \underbrace{-E_{z\sim Q_\phi(z|x)}\left[\log P_\theta(x|z)\right]}_{\text{Log likelihood}} - \underbrace{D_{KL}\left[\left(Q_\phi(z|x) \,\|\, P(z)\right)\right]}_{\text{Regularizer}}$$

reconstruction loss

regularizer



recognition model → $-E_{z\sim Q_\phi(z|x)}\left[\log\left(P_\theta(x|z)\right)\right]$

log likelihood

recognition model.

Bottleneck

generative model

takes the latent variable z and produces ⇒ the x

$P_\theta(x|z) = \mathcal{N}(\mu_\theta(z), \Sigma_\theta(z))$

We get a square error between the mean of the gaussian distribution and the data sample.!

Here

$$P_\theta(x|z) = \underbrace{\frac{1}{\sqrt{(2\pi)^k |\Sigma_\theta(z)|}}}_{\text{a multiplier}} \exp\left(\frac{(x-\mu_\theta(z))^T \Sigma_\theta^{-1}(z)(x-\mu_\theta(z))}{2}\right)$$

From the above we get

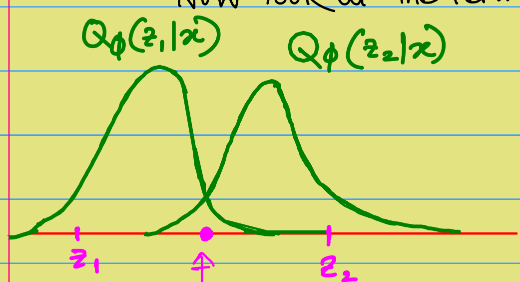
$$\log P_\theta(x|z) \propto (\underset{\text{data point}}{x} - \underset{\text{mean of the distribution.}}{\mu_\theta(z)})^T \Sigma_\theta^{-1}(z)(x-\mu_\theta(z))$$

New data $\hat{z} = \left(\frac{z_1+z_2}{2}\right)$ which is generated.

Now look at the term:

$$D_{KL}\left[\underbrace{Q_\phi(z|x)}_{\text{distribution of } z \text{ given } x} \,\|\, P_\theta(z)\right]$$

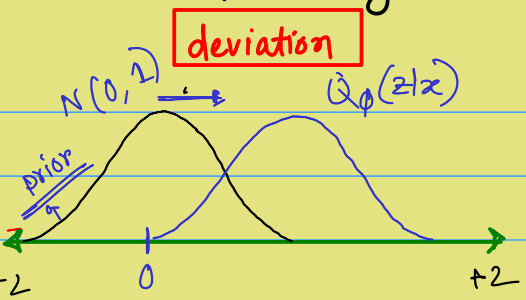
Here we are assuming that distribution of the $P_\theta(z)$ is a normal distribution with $N(0,1)$



weighted combination of the variables z_1 and z_2

→ We want our data (z) to be closer to a normal distribution. That is why divergence with $P_\theta(z) = N(0,1)$ is also calculated which is also working as a regularizer.

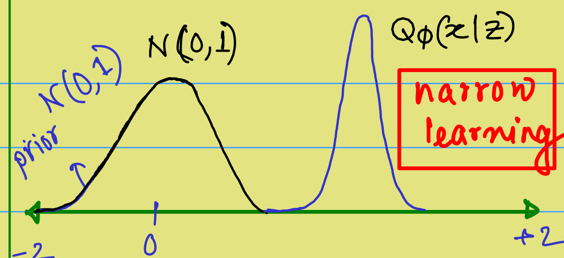
Data Fidelity with No KL divergence



deviates from the normal distribution. Will be of Normal shape but there is deviation.

(#) No term is there to bring it back to the normal distribution.

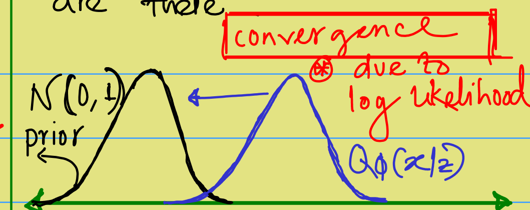
No data Fidelity But regularization.



(#) Network will cheat by learning a very narrow distribution. / No Large distribution.

Normal auto encoder !

Data Fidelity & Regularization both are there.



(*) shape equal to a $N(0,1)$ due to the regularized term KL divergence.

Aim and Objective: $D_{KL}\left(Q_\phi(z|x) \,\|\, P_\theta(z)\right) \longrightarrow$ Regularization Term.

easy choice of z with $N(0,1)$. We want $P_\theta(z)$ as close as possible to the $Q_\phi(z|x)$ so that they can be sampled easily.

Originally derived:

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

Now we put $\mu_2 = 0$ and $\Sigma_2 = 1$
and $\mu_1 = \mu_\phi(z)$ and $\Sigma_\phi(z) = \Sigma_2$ we get

$$D_{KL}\left(Q_\phi(z|z) \,\|\, P_\phi(x)\right)$$

$$= D_{KL}\left(N(\mu_\phi(z), \Sigma_\phi(z)) \,\|\, N(0,1)\right) =$$

$$\frac{1}{2}\left[tr(\Sigma(x)) + \mu_\phi(x)^T\mu_\phi(x) - K - \log|\Sigma_\phi(x)|\right]$$

= Here K is the dimension of the Gaussian

$tr(\Sigma(x))$ is the trace function. which is the sum of the diagonals $\Sigma_\phi(x)$.

Then determinant of a diagonal matrix is product of its diagonals.

above Equation becomes :

$$D_{KL}(N(\mu_\phi(x), \Sigma_\phi(x)) \| N(0, I))$$

$$= \frac{1}{2}\left(\sum_K \Sigma_\phi(x) + \sum_K \mu_\phi^2(x) - \sum_K 1 - \log \prod_K (\Sigma_\phi(x))\right).$$

$$= \frac{1}{2}\left(\sum_K \Sigma_\phi(x) + \sum_K \mu_\phi^2(x) - \sum_K 1 - \sum_K (\log \Sigma_\phi(x))\right)$$

$$= \frac{1}{2}\sum\left(\Sigma_\phi(x) + \mu_\phi^2(x) - 1 - \log \Sigma_\phi(x)\right).$$

For numerical stability we do the following like this

Model $\boxed{\Sigma_\phi(x) = e^{\Sigma_\phi(x)}}$

$$D_{KL}(N(\mu_\phi(x), \Sigma_\phi(x)) \| N(0, I))$$

$$= \frac{1}{2}\sum\left(\Sigma_\phi(x) + \mu_\theta^2(x) - 1 - \Sigma_\phi(x)\right)_{iv}$$