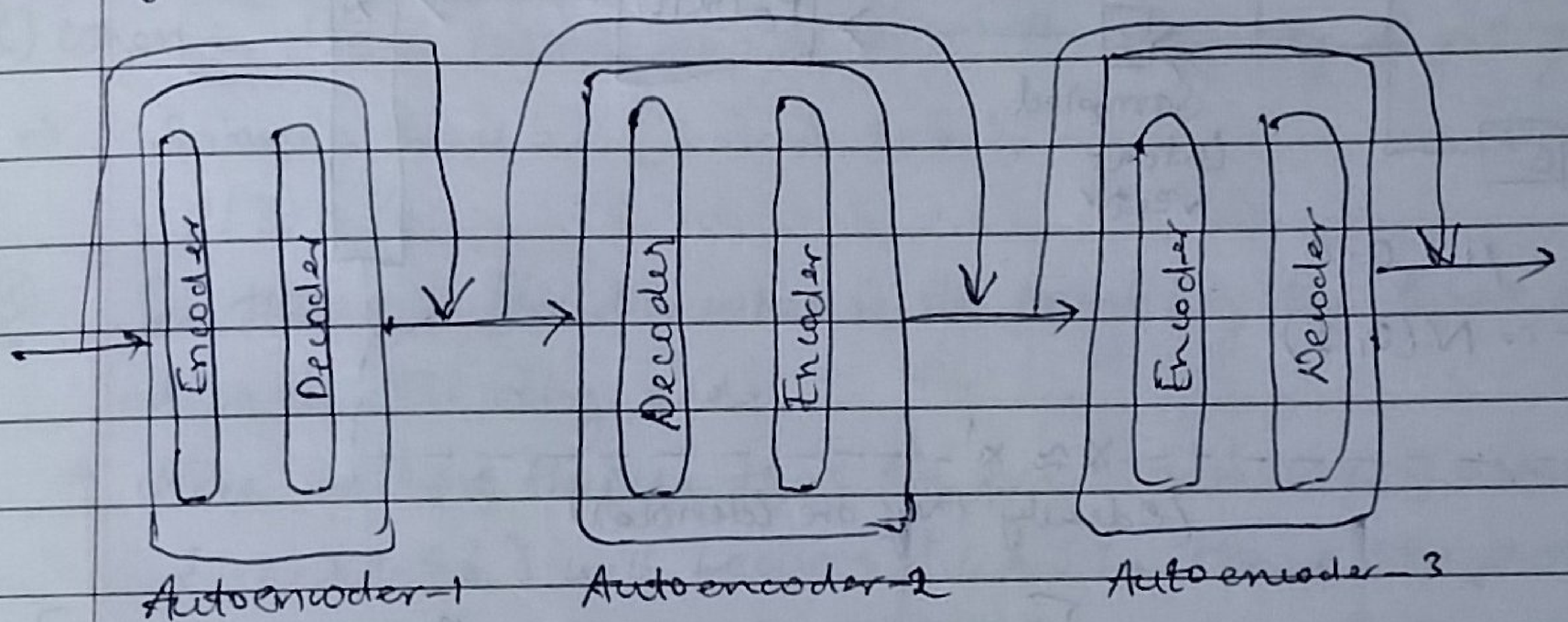
Variational Auto Encoders :→

We will first learn about the stacked autoencoder. As we know, that in some datasets have a complex relationship within the features. Thus, using only one Autoencoder is not sufficient. A single autoencoder might be unable to reduce the dimensionality of the input features. There fore, for such use cases, we use stacked autoencoders.

the stacked autoencoders are, as the name suggests, multiple encoders stacked on top of one another. A stacked auto encoder with three encoder stacked on top of each other is shown in the following figure



According to the architecture shown in the figure above, the input data is first given to auto encoder 1. The output of the auto encoder 1 and the input of the auto encoder 1 is then given as an input to auto encoder 2. Similarly, the output of auto encoder 2 and the input of auto encoder 2 are given as input to auto encoder 3. Thus, the length of the input vector for auto encoder 2. This technique also helps to solve the problem of insufficient data to some extent.

$$\text{Cost function} :\rightarrow \mathcal{L}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \left[x^{(i)} - f_\theta \left(g_\phi (x^{(i)}) \right) \right]^2$$

Lets learn about Denoising Autoencoders. → The purpose of DAE is to remove the noise.

DAE is primarily used for unsupervised learning. It is a variant of the standard auto encoder, but here is a critical twist :→ It is trained to reconstruct clean input data from corrupted (noisy) versions.

Simply, We deliberately corrupt the input data x to get a noisy version $\tilde{x}$, then train the network to reconstruct the original, clean data x from $\tilde{x}$

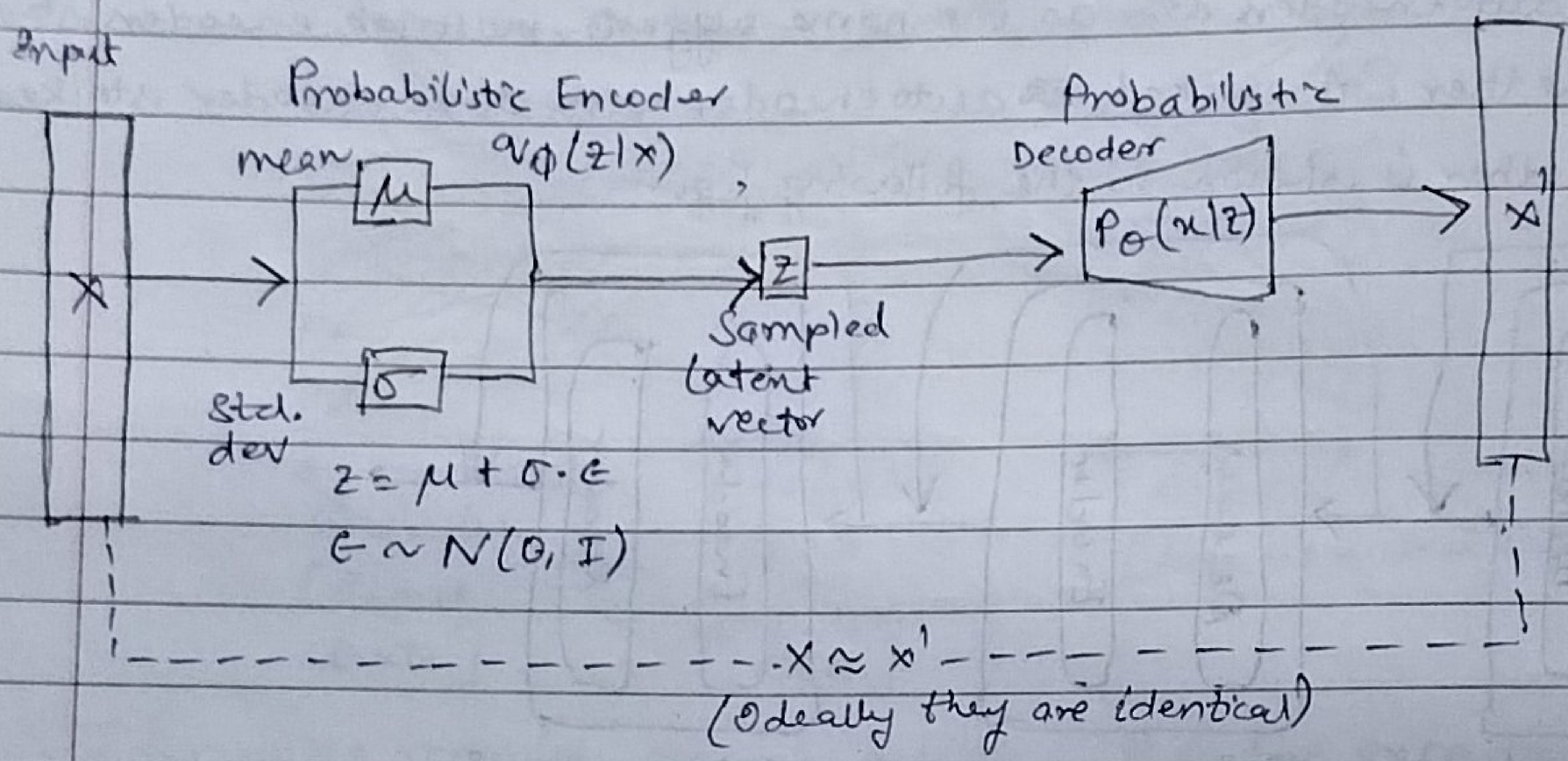
Why Noise → Preventing Overfitting, Encourage network to learn the more general features & Improves the performance

$$x^{(i)} \sim p_\theta(x^{(i)} | z^{(i)})$$

$$\text{Loss:-} \quad \mathcal{L}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \left[x^{(i)} - f_\theta(g_\phi(z^{(i)})) \right]^2$$

Page No.
Date

The bottle neck in the autoencoder contains the compressed version of the image that is input in a very optimised form that is the reduced dimension of the input image.



The only thing that is addedon are :→ [Probabilistic Encoder & Probabilistic Decoder]

Prerequisite.

Probability →
- $p(x)$ ⇝ probability
- $p(x|y)$ ⇝ conditional probability
- $E[\cdot]$ ⇝ Expectation
- KL Divergence

- $p(x)$: defines the probability of random variable x [$p(x) \leftarrow x$]
- $p(x|y)$: defines as the probability of a random variable X provided Y has happened, it also called conditional probability.

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)} \quad \text{(Likelihood, prior probability)}$$

$p(y|x)$ ↙ Posterior probability

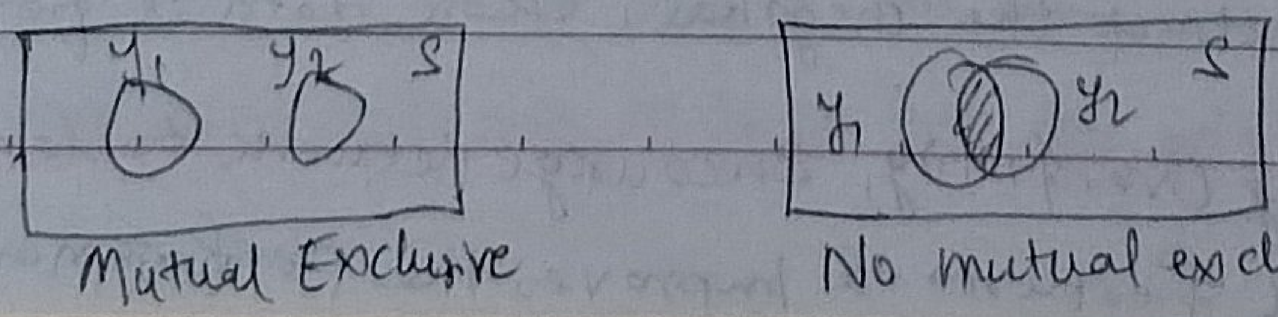
⇝ Bayes theorem

$$\rightarrow \frac{p(x, y)}{p(y)} \text{ ⇝ Joint distribution} \quad \text{— (1)}$$

(*) Theorem of total probability :→

Let $y_1, y_2, y_3 \ldots y_N$ be a set of mutually exclusive events (i.e. $y_i \cap y_j = \emptyset$) & event X is the union of N mutually exclusive events, then

$$p(x) = \sum_{i=1}^{N} p(x/y_i)\, p(y_i) \quad \text{— (2)}$$



from (2) in (1) $\left\{ P(Y|X) = \frac{P(X|Y)\ P(Y)}{\sum_{i=1}^{n} P(X|Y_i)\ P(Y_i)} \right\}$

Page No.
Date

(*) Expectation of a Random Variable X i.e. E(X)

Expected value of random variable is a weighted average of the possible values of X can take, each value being watched according to the probability of that event defined as :→

$$E(X) = \sum_{i=1}^{k} x_i\, P(x = x_i) \simeq E_P(X)$$

(Q) When a die is tossed one. what is the Probabity of getting 3

→ Sample Space = {1, 2, 3, 4, 5, 6}.
$P(3) = 1/6$

(Q) In tossing a fair die, what is the probability the 3 has occured conditioned on the toss being odd.

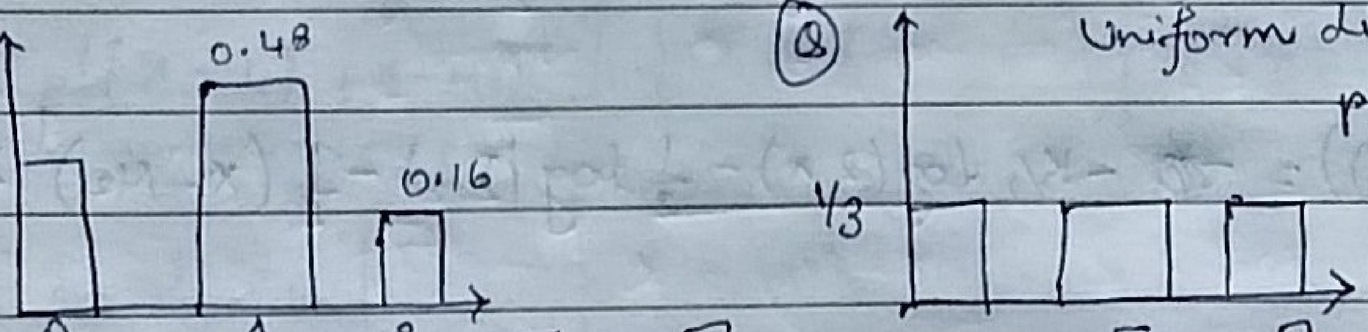
⇒ Since we have the idea that the question is saying about the odd so, {1, 2, 3, 4, 5, 6} will become {1, 3, 5}. Hence the probability will be reduced to 1/3.

(#) KL Divergence :- (Kullback - Leibler divergence) → It is measure of how one probability distribution is different from the second. For the discrete probability distribution P & Q, the KL divergence b/w P & Q is defined as

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

$$= \sum_{x} P(x)\ \log\left(P(x)/Q(x)\right)$$

Examples :- (P) (Q) Uniform distribution with $P = 1/3$



[P, Q ~ 0] [P, Q ~ 1] [P, Q ~ 2]

$$D_{KL}(Q||P) = \frac{1}{3}\ln\left(0.333/0.36\right) + \frac{1}{3}\ln\left(0.333/0.48\right) + \frac{1}{3}\ln\left(0.333/0.16\right)$$

$$= 0.09637 \text{ nats} \rightsquigarrow \text{unit}$$

Properties

① $KL(P||Q)$ or $KL(Q||P) \geq 0$

② $KL(P||Q) \neq KL(Q||P)$ {Not Symmetric}

Page No.
Date

Suppose we have two multivariate normal distributions defined as :→

$$P(x) = N(x; \mu_1, \Sigma_1)$$
$$Q(x) = N(x; \mu_2, \Sigma_2)$$

[N ⇒ Normal distributions]

where, μ_1 & μ_2 are the means & Σ_1 & Σ_2 are the covariance matrix.

Recall, the multivariate normal density is defined as :→

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

→ if the two distributions have the same dimension 'K'.

$$D_{KL}(P(x)||q(x)) = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - d + tr(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1)$$

Derivation :→ We know, $KL(P(x)||Q(x)) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$ —①

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

⇒ Taking log both sides → ③

$$\log(P(x)) = -K/2 \log(2\pi) - \frac{1}{2} \log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) - ③$$

Similarly $\log(Q(x)) = -K/2 \log(2\pi) - \frac{1}{2} \log|\Sigma_2| - \frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1} (x-\mu_2) - ④$

We can write eqn ① as :→

$$KL(P(x)||Q(x)) = \sum_x P(x)(\log(P(x)) - \log(Q(x))) - ①$$

Substituting ③ & ④ in ① results in :→

$$\Rightarrow KL(P(x)\|Q(x)) = \sum_x P(x)\left\{-\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-M_1)^T\Sigma_1^{-1}(x-M_1)\right.$$

$$\left. + \frac{K}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_2| + \frac{1}{2}(x-M_2)^T\Sigma_2^{-1}(x-M_2)\right\}$$

$$\Rightarrow KL(P(x)\|Q(x)) = \sum_x P(x)\left\{+\frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_2|} + \frac{1}{2}(x-M_2)^T\Sigma_2^{-1}(x-M_2)\right.$$

$$\left. - \frac{1}{2}(x-M_1)^T\Sigma_1^{-1}(x-M_1)\right\} \text{ — (3)}$$

Now, lets consider part by part.

$$\sum_x P(x)\frac{1}{2}(x-M_1)^T\Sigma^{-1}(x-M_1) = E_P\left[\frac{1}{2}(x-M_1)^T\Sigma^{-1}(x-M_1)\right]$$

↑ we can write this by the Rule of the Expectation

- FROM LINEAR ALGEBRA

$\rightarrow E(x^TAx) = E(tr(x^TAx))$ — (a)

$= E(tr(Axx^T))$ — (c)

$= tr(E(Axx^T))$ — (e)

Lets rewrite again,

$\Rightarrow \frac{1}{2}E_P[(x-M_1)^T\Sigma_1^{-1}(x-M_1)]$

$E_P[tr(\frac{1}{2}(x-M_1)^T\Sigma_1^{-1}(x-M_1)]$

$E_P[tr(\frac{1}{2}(x-M_1)(x-M_1)^T\Sigma_1^{-1}]$ — (d)

$\rightarrow tr[E_P(\frac{1}{2}(x-M_1)(x-M_1)^T\Sigma_1^{-1})]$ — (e)

$\rightarrow tr[E_P[(x-M_1)(x-M_1)^T]\frac{1}{2}\Sigma_1^{-1}]$

$E_P[(x-M_1)(x-M_1)^T]$ → Covariance matrix

$\Rightarrow tr[\Sigma_1\frac{1}{2}\Sigma_1^{-1}]$

$\Rightarrow tr[I_K] = K$ — (6)

$\therefore \boxed{E_P\left[\frac{1}{2}(x-M_1)^T\Sigma^{-1}(x-M_1)\right] = K}$ → (7)

Trace & Expectation trick

→ if x is scalar then $E(x) = E(tr(x))$ since (a)

· trace of x is scalar

→ $tr(AB) = tr(BA)$ — (b)

→ $tr(ABC) = tr(BCA)$

$= tr(CAB)$ — (c)

→ $tr(ABC) \neq tr(ACB)$ — (d)

→ $E(tr(x)) = tr(E(x))$ — (e)

↳ IMP

trace: → sum of all diagonal elements in a matrix

Now, consider the second part

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

$$\rightarrow (A+B)^T \Sigma_2^{-1}(A+B)$$

$$\rightarrow (A^T+B^T)\Sigma_2^{-1}(A+B) =$$

Data

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

↳ {Adding μ_1 & substracting μ_1}

$$\rightarrow \sum_x p(x)\Big[\underbrace{\tfrac{1}{2}(x-\mu_1)^T\Sigma_2^{-1}(x-\mu_1)}_{A^T\Sigma_2^{-1}A} + \underbrace{\tfrac{2}{2}(x-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2)}_{\{B^T\Sigma_2^{-1}A,\ A^T\Sigma_2^{-1}B\}} + \underbrace{\tfrac{1}{2}(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)}_{B^T\Sigma_2^{-1}B}\Big]$$

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

Expanding we get,

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

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

constant

$$\rightarrow \left\{tr\left\{\frac{\Sigma_2^{-1}\Sigma_1}{2}\right\}\right\} + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) + \underbrace{0}_{\text{proved (next)}} \quad \text{— (8)}$$

↓ similar to earlier derivation

$E[\text{constant}] = \text{constant}$

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

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

$$\rightarrow (\mu_1-\mu_1)^T\Sigma_2^{-1}(\mu_1-\mu_2) = 0 \quad [\text{Proved}]$$

(*) $\boxed{E_{x\sim p(x)}\ \tfrac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) = k}$ — (7)

$\boxed{E_{x\sim p(x)}\ \tfrac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) = tr\{\Sigma_2^{-1}\Sigma_1\} + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)}$ ↳ (8)

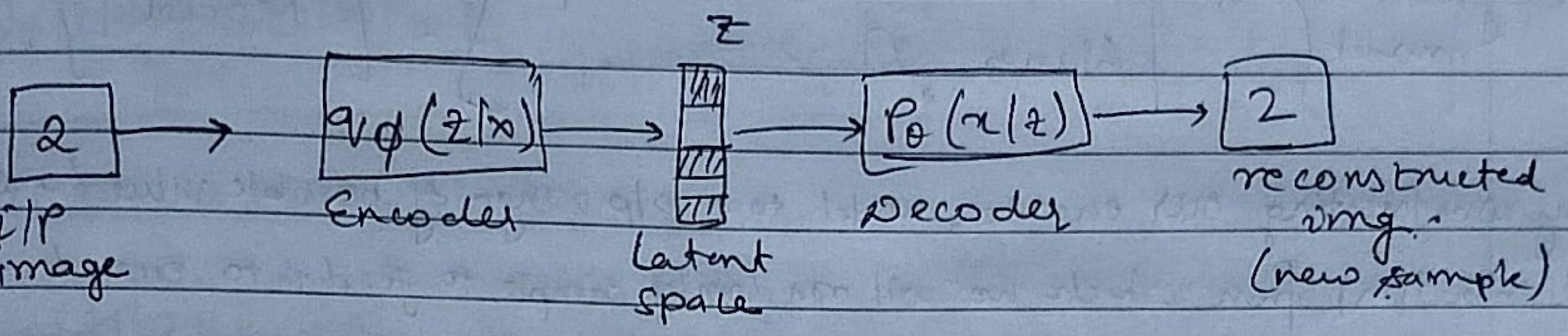
from (7) & (8) in (5)

$$\boxed{KL(P(x)\|Q(x)) = \frac{1}{2}\log\left|\frac{\Sigma_1}{\Sigma_2}\right| + k + tr\{\Sigma_2^{-1}\Sigma_1\} + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)}$$

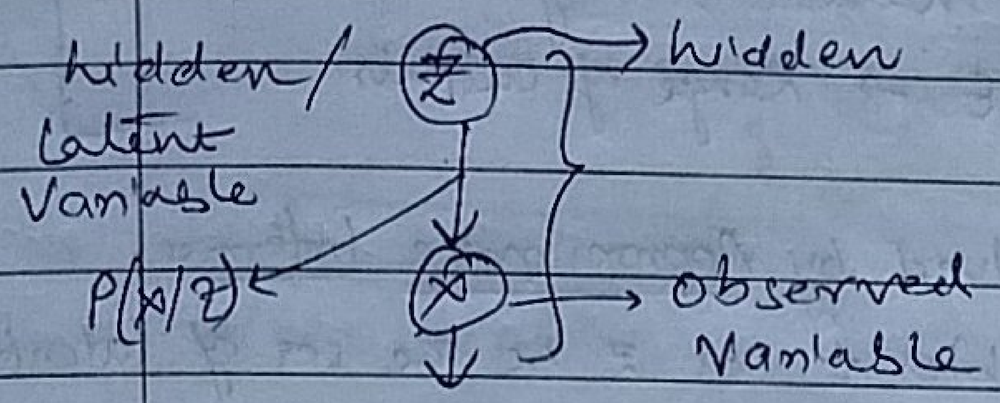
(Proved)

The goal of VAE :->

The goal of VAE is to find a distribution $q_\phi(z|x)$ of some latent variables which we can sample from $z \sim q_\phi(z|x)$ to generate new samples $x' \sim p_\theta(x|z)$

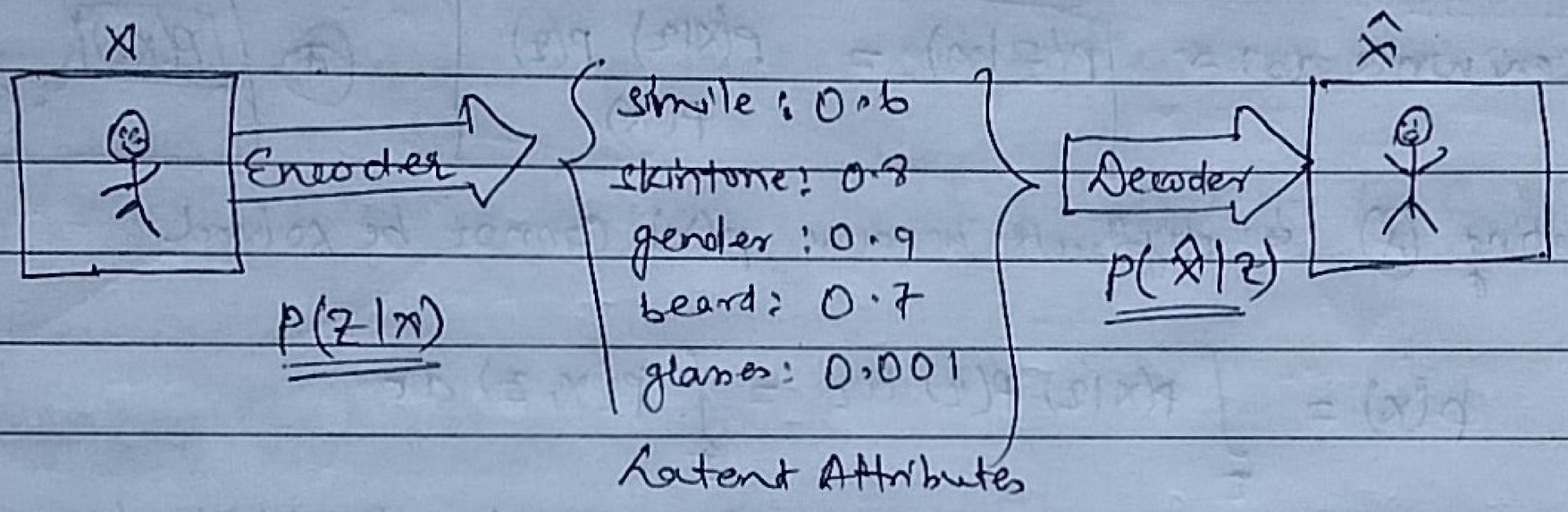


- Latent Variable (z) :- We model system as a collection of random variables.



The edge drawn from $z \to x$ is the conditional distribution $p(x|z)$

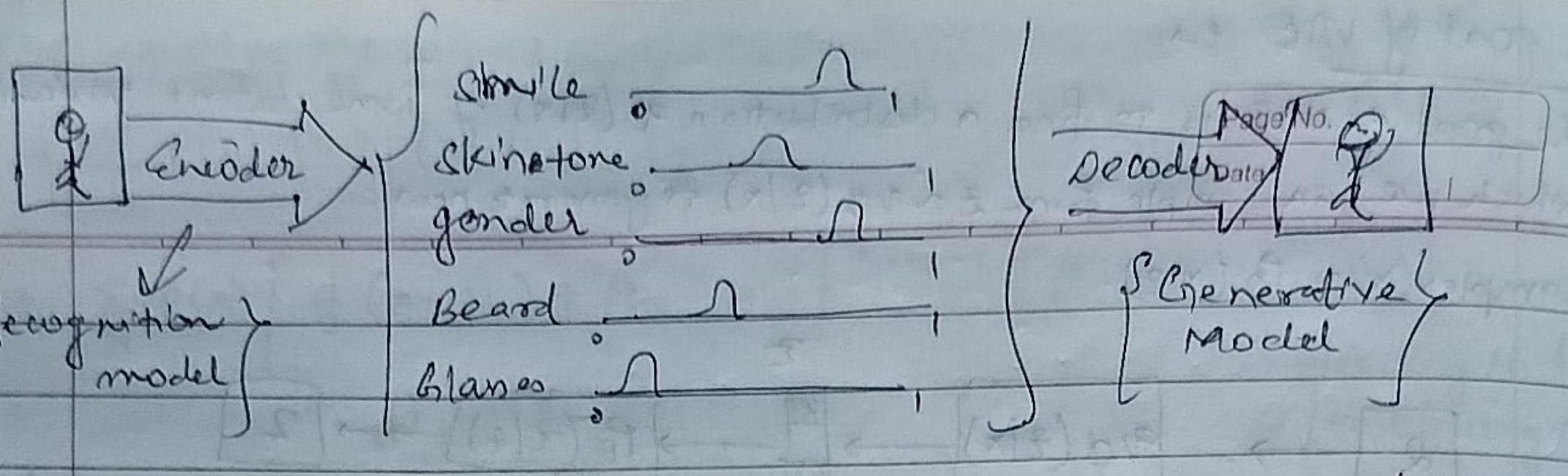
Latent Variables corresponding to a real feature of the object that have not been measured (may be technology is not available to do that)



→ In the above examples, we trained autoencoder on a large dataset of faces with encoding dimension of 6. An ideal autoencoder will learn the descriptive attributes of faces such as skin color, smile etc. On order to describe observation in some compressed form

→ In above example, we have described the i/p image in terms of latent variables using single value to describe each attribute. For instance, what single value you will assign for photo of Monalisa?

→ With the approach of PD, we can represent each latent attribute for a given i/p as a probability distribution. When decoding, we will randomly sample from each latent state distribution to generate a vector as i/p for an decoder model.



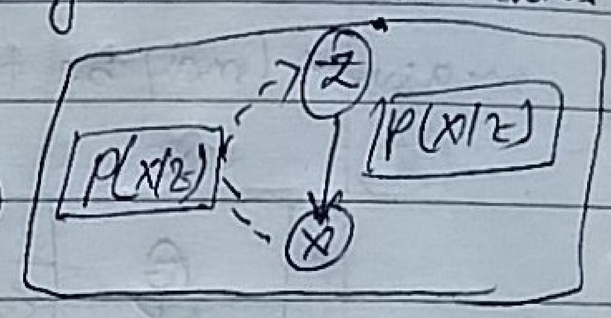
By constructing our encoder model to a o/p range of possible value (a statistically distribution) from which we will randomly sample to feed it to our decoder model. The values which are nearby each other in latent space must corresponds to similar reconstruction

Auto Encoder → Only one output
VAE → Range of outputs

- The derivation of loss function can be solved by Approximate Inference.

Let x be a set of observed variables & let z be the set of latent variables with joint distribution $p(z,x)$. Then the inference problem is to compute the conditional distribution of latent variables given the observations i.e. $p(z/x)$

we can write as: → $$p(z|x) = \frac{p(x|z)\,p(z)}{p(x)}$$ — (A)



Evaluating (A) is difficult because $p(x)$ cannot be solved

Reason:-

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

This integral is not available in closed form or is intractable (i.e. requires exponential time to compute) due to multiple integrals involved for latent variable vector z. → we have 6 variable so we need 6 integrals

Alternative: the alternative is to approximate $p(z|x)$ by another distribution $q(z|x)$ which is defined in such a way that it has tractable solution. This is done via variational inference. the main idea of VI is to pose the inference problem as an optimization problem How?

By modelling $p(z|x)$ using $q(z|x)$ (where $q(z|x)$ has a simple distribution such as Gaussian)

as discussed let us calculate KL b/w $p(z/x)$ & $q(z/x)$

$$D_{KL}\left(Q_\phi(z|x) \,\|\, P_\theta(z|x)\right) = \sum_z Q_\phi(z|x) \log\left(\frac{Q_\phi(z|x)}{P_\theta(z|x)}\right)$$

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

Page No.
Date

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

Putting A to B (Substituting)

~~$D_{KL}\left(Q_\phi(z|x) \,\|\, P_\theta(z|x)\right) = \sum_z Q_\phi(z|x) \log\left(\frac{Q_\phi(z|x)}{P_\theta(z|x)}\right)$~~

~~$= E_{z\sim Q_\phi(z|x)} \log\left(\frac{Q_\phi(z|x)}{P_\theta(z|x)}\right)$~~

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

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

Since the expectation is over z & $P(x)$ doesn't involve z, it can be moved out.

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

∴ Rearranging the Equations:-

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

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

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

This is the VAE objective function, where the first term represents the reconstruction liklihood & the second term ensures that our learned distribution Q is the similar to the prior distribution P.

$$E_z \log \frac{Q_\phi(z|x)}{P_\theta(z)}$$

Loss function = − Objective function

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

Proved!!!

Page No.
Date

also, $\underbrace{\log P_\theta(x)}_{\text{Reconstruction}} - \underbrace{\left[D_{KL}\, Q_\phi(z|x) \,||\, P_\theta(z|x)\right]}_{\text{Regularizer}} = -\mathcal{L}(\theta,\phi)$

or

Now, our target is to find optimal θ, ϕ such that $\left\{\theta^*, \phi^* = \underset{\theta,\phi}{\operatorname{argmin}}\left(\mathcal{L}(\theta,\phi)\right)\right\}$

- Understanding each Variable

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

Inputs — log likelihood — Regularizer

$Q_\phi(z|x)$ → Recognition model mapping $x \to z$

z → Bottleneck

$P_\theta(z|x)$ → Generative model mapping z to x

$P_\theta(x|z) = N\left(\mu_\theta(z), \Sigma_\theta(z)\right)$

So, when we take log of Gaussian we get a square error b/w the data sample x & mean of the Gaussian distribution

Role

$Q_\phi(z_1|x)$ $Q_\phi(z_2|x)$

z_1 z_2

Latent Variable space

KL divergence not allows pdf of latent variables not collapse with 0 variance but penalize if deviates from $N(0,1) = P_\theta(z)$

For multi gaussian env :-

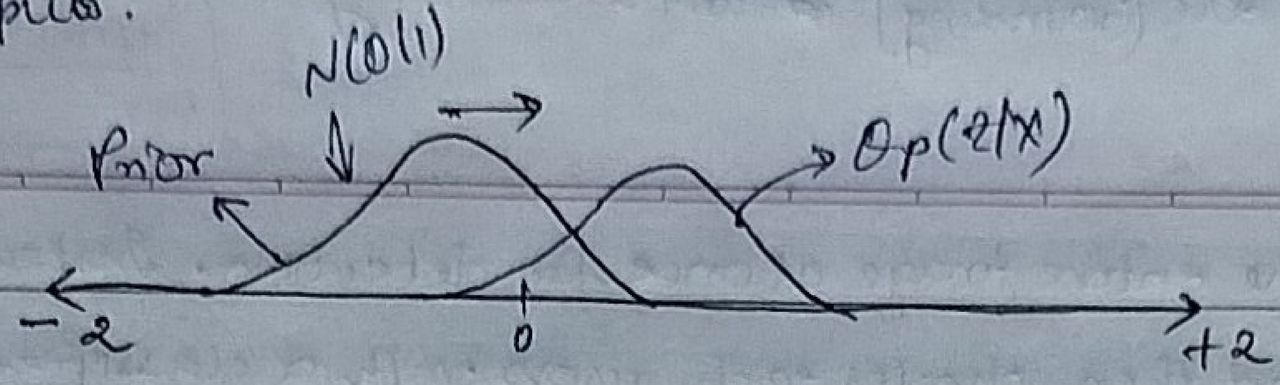
$$P_\theta(x|z) = \frac{1}{\sqrt{(2\pi)^k \left|\Sigma_\theta(z)\right|}} \exp\left[\frac{x-\mu_\theta(z)^T \Sigma_\theta^{-1}(z)\,(x-\mu_\theta(z))}{2}\right]$$

or $\log P_\theta(x|z) \propto \left[(\hat{x}-\mu_\theta(z))^T \Sigma_\theta^{-1}(z)\,(x-\mu_\theta(z))\right]$

Squared Reconstruction error → data fidelity in inverse problem

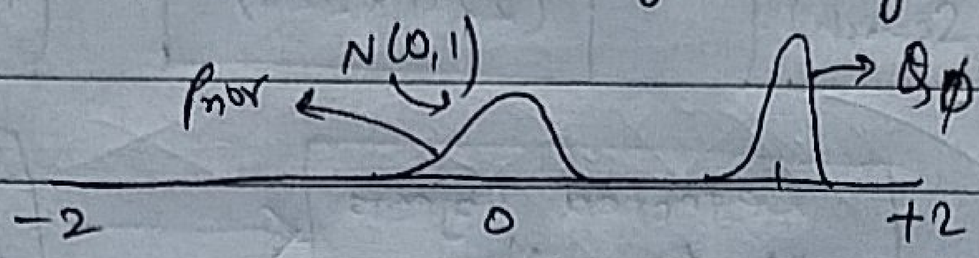
$\mathcal{L}$ = data fidelity term + KL div term

① Penalizing reconstruction loss (without data fidelity):- encourages distribution to describe inputs.



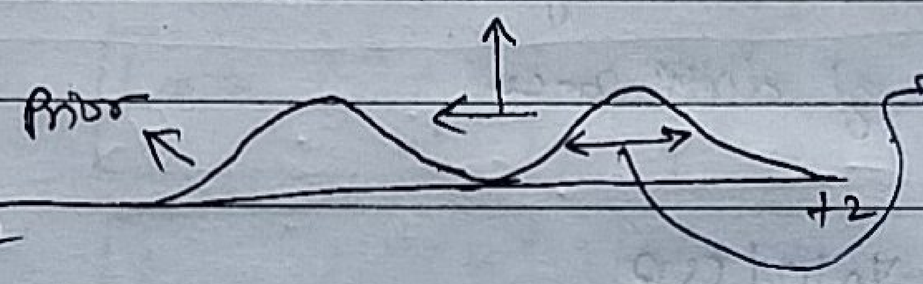
Our distribution deviates from the prior to describe some characteristics of data

② Without Regularization networks cheats by learning narrow distributions



with the small variance, this distribution is eff rep single value.

③ Attraction blw the two distribution is due to 1st term



Sufficient Variance is ensured using KL divergence

- ORIGINAL → $N(0,1)$
- OUR DISTRIBUTION → $N(\mu_\phi(x), \Sigma_\phi(x))$

Final loss (NO derivation):-

$$\mathcal{L}(\theta,\phi) = -E_{z}\left[\log P_\theta(x|z)\right] + \frac{1}{2} E_K\left[\exp(\Sigma_\phi(x)) + \mu_\phi^2(x) - 1 - \Sigma\phi(x)\right]$$

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

Optimised loss:-

$$\hat{\phi}_i = -\frac{1}{S}\sum_{s=1}^{S} \nabla_\phi \left[\log P_\theta(x|z^{(l)})\right]$$

where, $z^{(l)} = \mu_\phi(x) + \varepsilon \odot \sigma_\phi(x)$

$\varepsilon^{(l)} \sim N(0,1)$