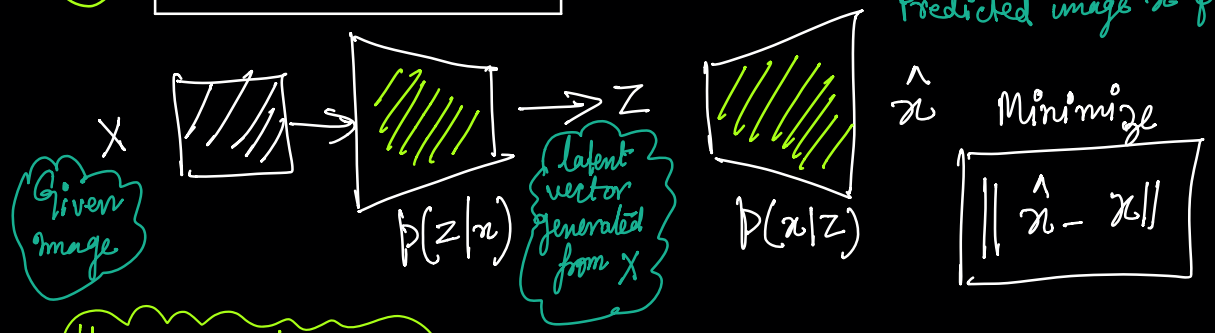
Probability of (z) given (x)

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

) Can be converted into ② [Bottleneck]

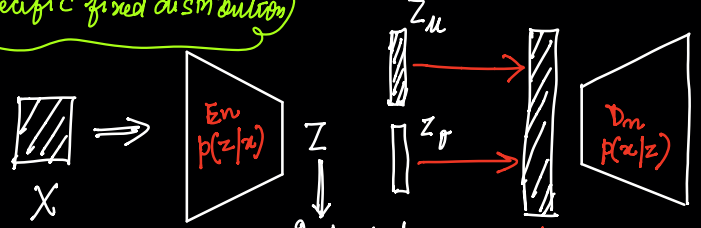
(A) Basic Auto encoder



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 Auto encoders



Instead of the ② we want our encoder to get μ, σ i.e. z_μ, z_σ

Sampling layer: Sample a point from $G(z_\mu, z_\sigma)$

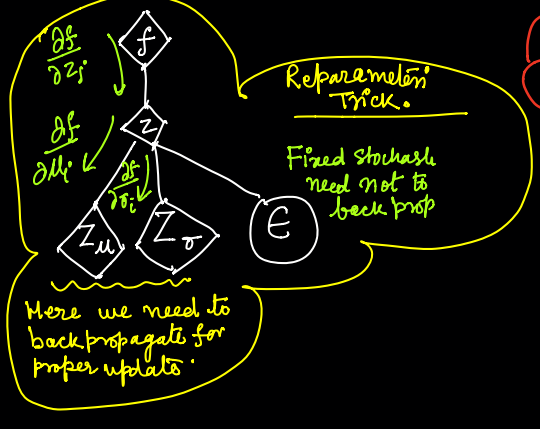
Losses

a) Reconstruction loss: $\min \|x - \hat{x}\|$

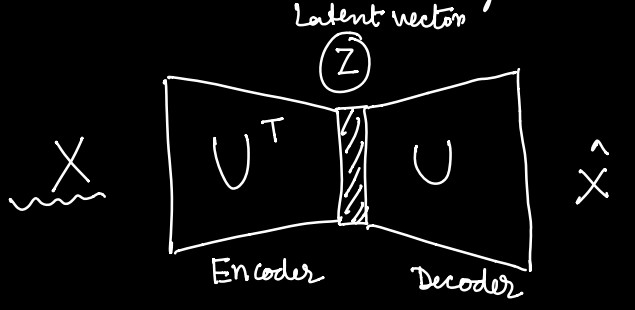
b) KL-Div($G(z_\mu, z_\sigma)$, $N(0,1)$)

$(n \times 1)$ $(n \times n)$ ⇒ take only diagonal (variance)

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



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.)