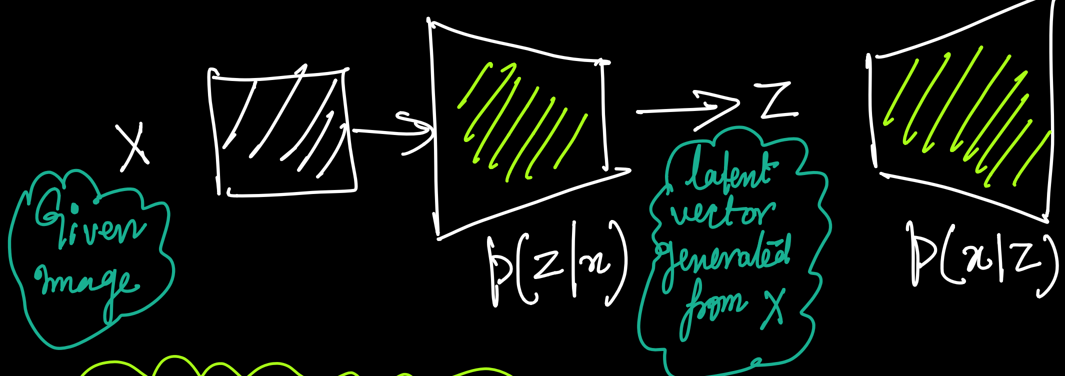
Probability of z given x

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

x can be converted into z [Bottleneck]

A Basic Autoencoder

Predicted image $\hat{x}$ from z



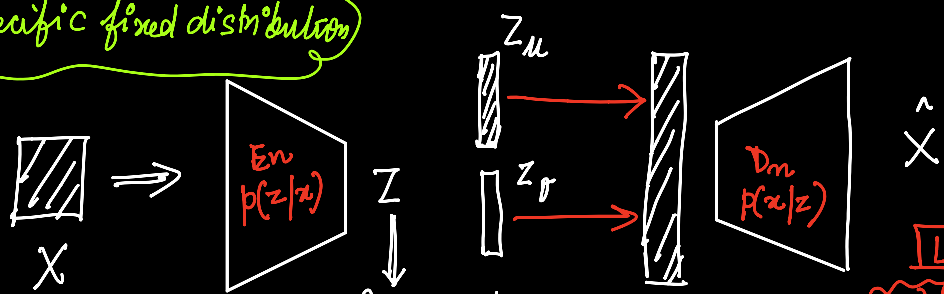
Minimize

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

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

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

B Variational Autoencoders



Instead of the Z we want our encoder to get μ, σ Z_μ, Z_σ

Sampling layer

Sample a point from $G(Z_\mu, Z_\sigma)$

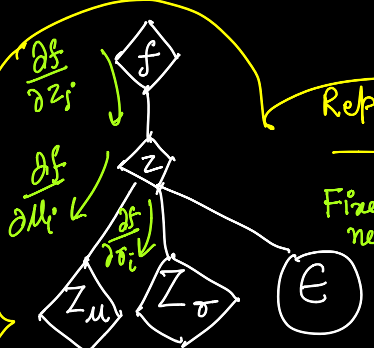
$(n \times 1)$ $(n \times n)$ ⇒ take only diagonals (variance)

Losses

a) Reconstruction loss

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

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



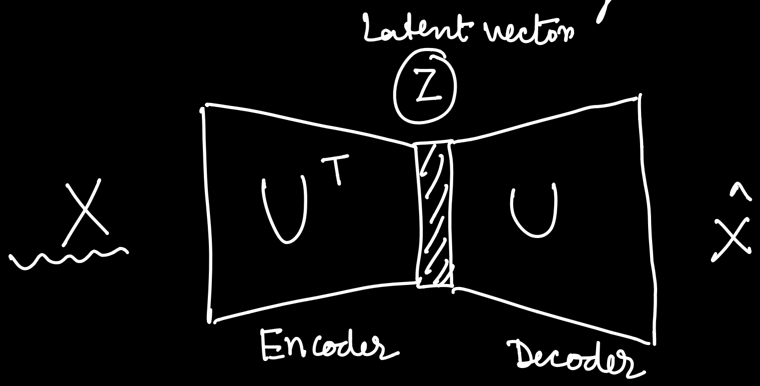
Reparameterization Trick.

Fixed stochastic need not to back prop

Here we need to back propagate for proper update.

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

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 donot ensure that Such A.E and PCA both learns the identical basis but may span the Similar space

Encoder (E)

(1) $\underbrace{Z}_{p\times 1} = \underbrace{U^T}_{p\times d} \underbrace{X}_{d\times 1}$

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

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

Encoder is learning some transformation that can convert $\underbrace{X}_{i/p}$ to $\underbrace{Z}_{\text{latent vector}}$

Decoder (D)

(2) $\underbrace{\hat{X}}_{d\times 1} = \underbrace{U}_{d\times p} \underbrace{Z}_{p\times 1}$

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

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

In A.E $\boxed{U}$ may not be learned as orthonormal.

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

Important Concepts

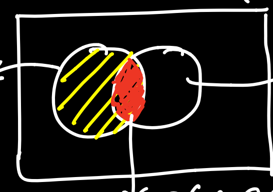
A) Bayes Theorum:

(i) Conditional probability ÷ events are (A) & (B)

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

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

P(B)



P(A)

$P(A \cap B) = P(B \cap A)$

$$\Rightarrow P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \Rightarrow \boxed{\underbrace{P(A|B)}_{\text{Hypothesis}} = \frac{P(B|A) \cdot \overbrace{P(A)}^{\text{Prior}}}{P(B)}}$$

Evidence

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:

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

X Event	$p(x)$ Probability	$-\log(p(x))$ Information
(1) Virat Scored a centure.	↑ (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]

So basically rare events carry more information

C Average of Information is Entropy (H):

The expected value of information w.r.t any event (X) averaged over all values (X) can attain is Entropy (H).

This is the expected value of $[\log p(x)]$ wrt $p(x)$

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

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)

Amount of information in (q) distribution

Amount of information in (p) distribution

$$\underbrace{-\sum q(x) \log q(x)}_{\text{This expectation wrt } q(x)} + \underbrace{\sum p(x) \log p(x)}_{\text{This expectation is wrt } p(x)}$$

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

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

$\Downarrow$ This is the Cross entropy between (p) and (q) distributions.

$\Downarrow$ This is (-ve) entropy of (p) distribution

Now 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 $p(x)$.

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

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

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

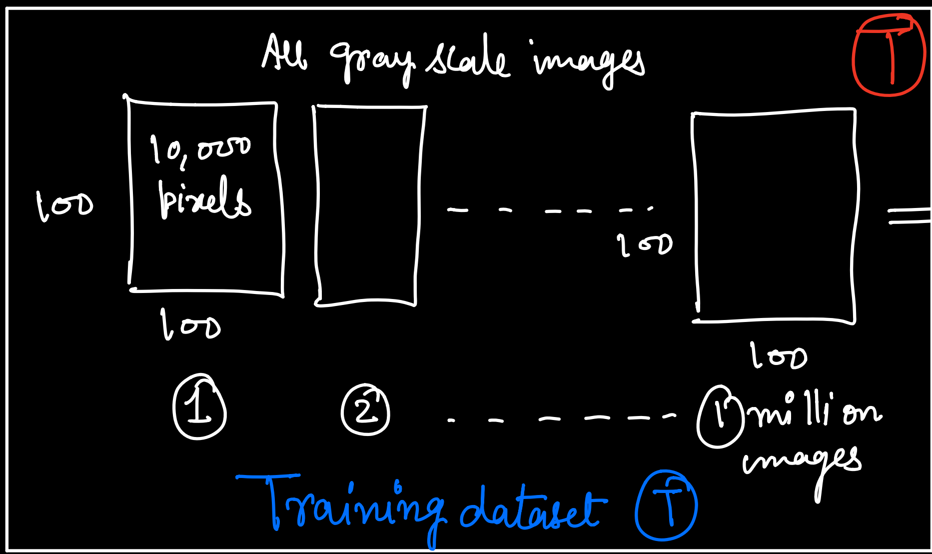
⊛ KL-Div is not Symmetric as KL-Div$(p|q) \neq$ KL-Div$(q|p)$

↳ Hence it is a distance measure b/w a divergence.

⊛ KL-Div ≥ 0 it is always +ve and b/w (0 & 1)

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

(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-occurences patterns.

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

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

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

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 $[\theta]$.

$$\boxed{\theta^* = \underset{\theta}{\arg\max} \left[P(x \in T) \right]}$$

minimizing the rest.

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

- (a) <u>Classification</u>: Helps us to discriminate b/w images that are coming from (T) or rest.
- (b) <u>Generative modeling</u>: 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.

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

- 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.

latent variable

$$Z_1, Z_2 \cdots Z_{100} \xrightarrow{\text{Influences}} X_1, X_2, X_3 \cdots\cdots X_{10,000}$$

observation

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 - - - .

external Parameters

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

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