Continuation of VAE- theory

02/03/25

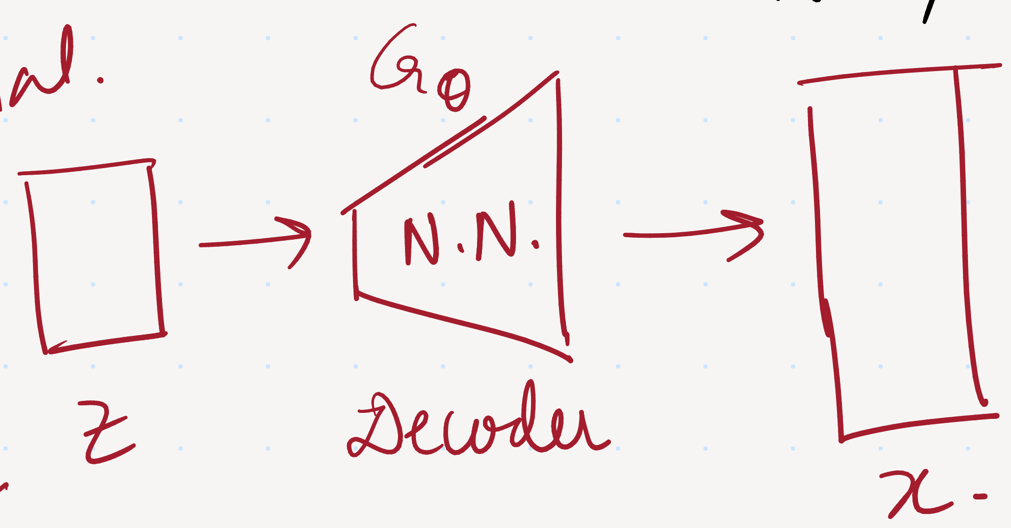
(A more rigorous treatment of the topic)

Sample an image from a distribution
Training data for Neural Networks.
} need for Generation of images.

Generative Model

θ = parameters of the N.N/Generator.

Sampled from a standard Normal R.V.



$x = G_\theta(z)$.

Random

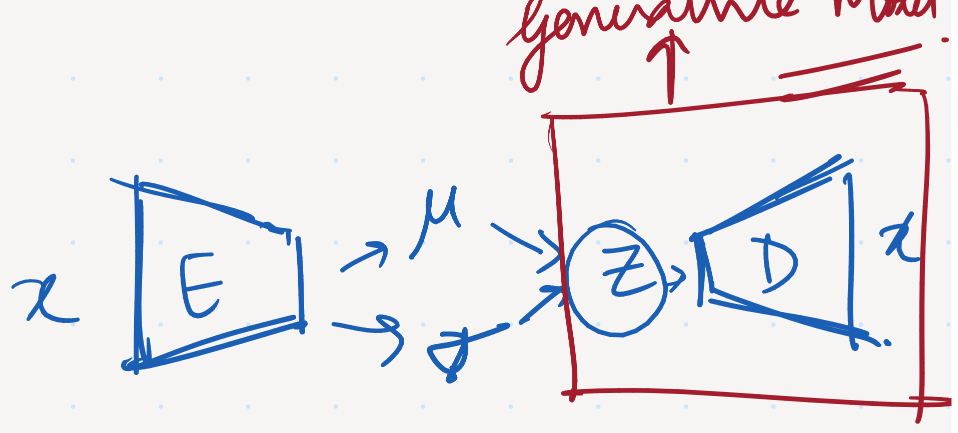
Deterministic.

Induced Random Variable.

Latent vector/code/variable/representation.

Sample → high dimensional images.

Typically, $z \sim \mathcal{N}(0, I)$ → low dimensional.



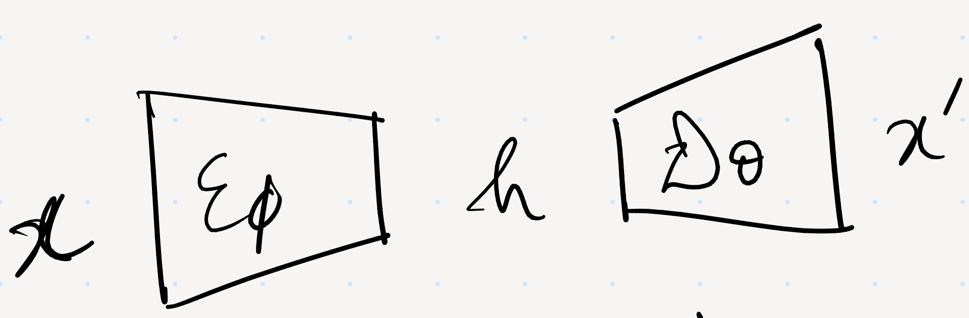
$\mu \, \varepsilon + \sigma$ —

$\varepsilon \sim G(0, I)$.

V.A.E.

Scale with mean & add variance for reparam trick.

Autoencoders :-



L2-reconstruction error.

$$x' = D_\theta(E_\phi(x)) \simeq x.$$

$$\min_{\theta, \phi} \| x_i - D_\theta(E_\phi(x)) \|^2$$

Autoencoder is not a generative model, since it is not sampling from a data distribution. Generative model → Define a distribution.

Training a low latent-dimensional generative model by likelihood.

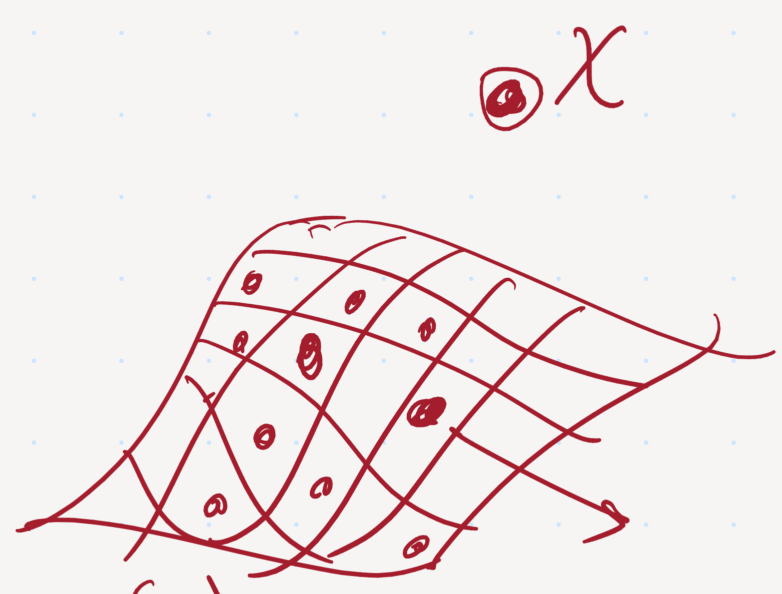
Given data $\{x_i\}_{i=1,2,\dots n}$ train a generative model to maximize the likelihood of the observed data under our model.

If Generative Model :-

$G_\theta : \mathbb{R}^k \to \mathbb{R}^d$, w/ $k<d$.

(low dim) $z \longrightarrow x$ (high dim).

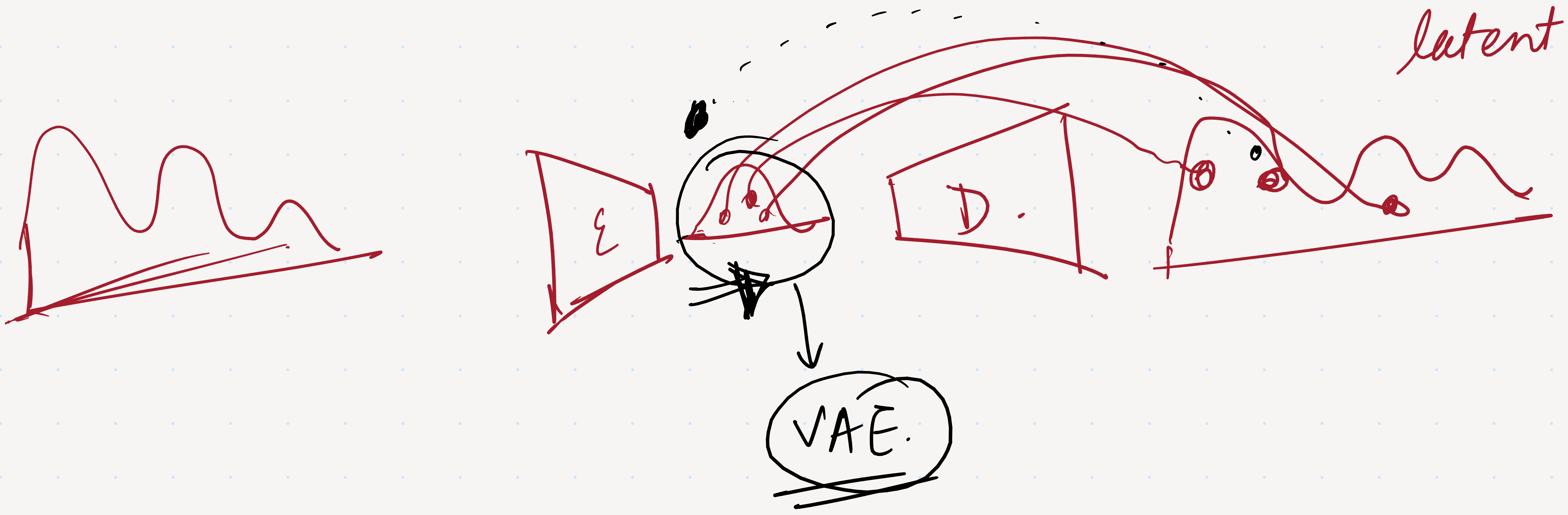
then $p(x) = 0$ almost everywhere



Range(G). non-zero likelihood.

↓

manifold of the latent z.



∴ we have a non-zero likelihood only on the lower dimensional subset of space (subspace) (i.e., Range(G))

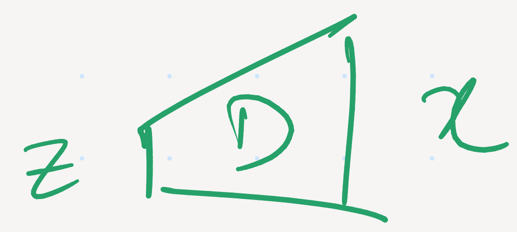
If we pick a randomly generated point x, off the manifold of G, then this point won't have any probability, i.e., on the higher dimension.

So, we can't directly optimize likelihood.

To have a non-zero likelihood everywhere, define noisy observation model. / Noisy inference model.

$$\boxed{p_\theta(x|z) = \mathcal{N}(x\,;\, \underbrace{G_\theta(z)}_{\mu}\,,\, \underbrace{\eta I}_{\sigma})} \rightarrow \text{likelihood.}$$

$z \rightarrow$ induce a distribution in image space governed by θ. Under a simple prior $p(z)$, this induces a joint distribution $p_\theta(x,z)$.



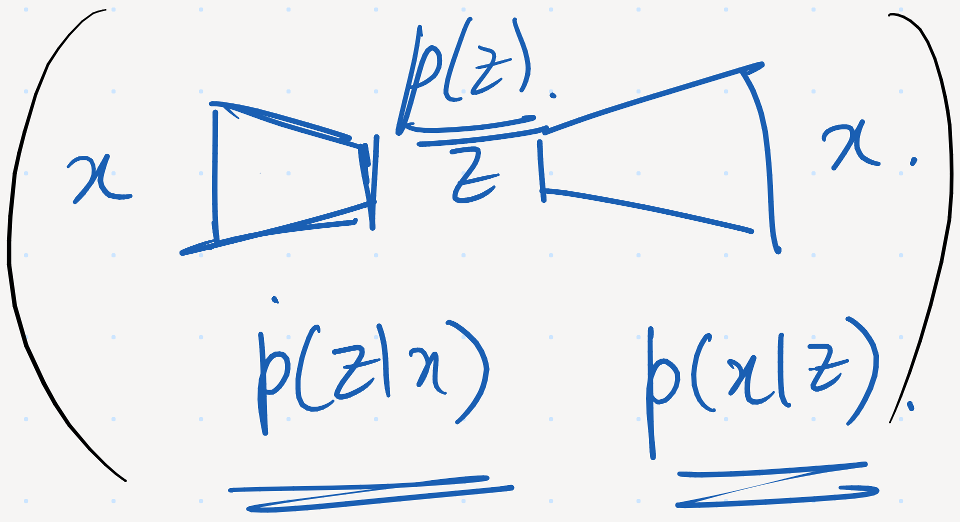
$p(x|z)$.

∘ intractable to evaluate at each iteration, hence we optimise a lower bound instead.

$$p(x) = \int p(x|z) \cdot p(z)\, dz.$$

likelihood of z

likelihood of $x|z$

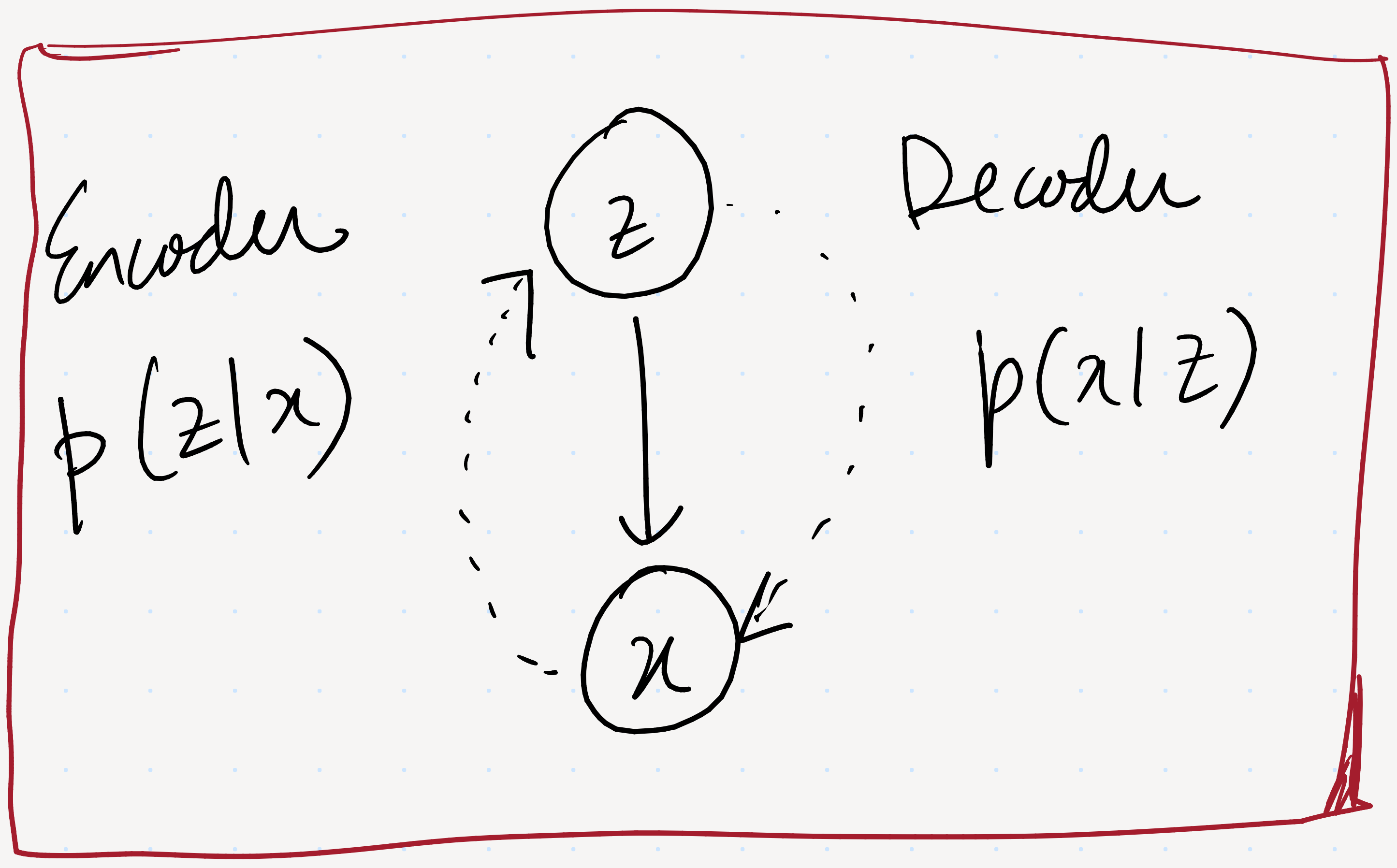


Maximize the log-likelihood of the data.

likelihood prior

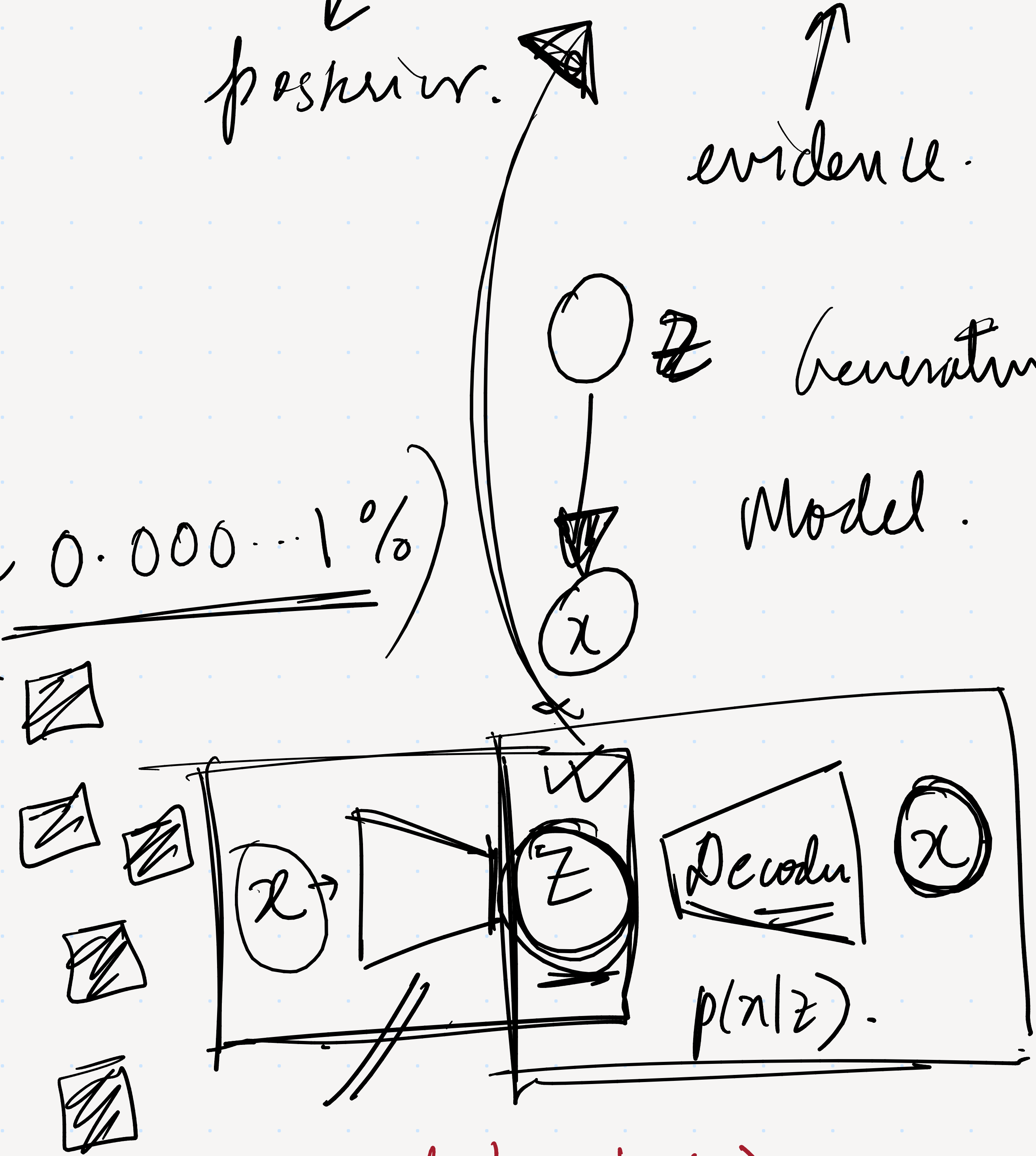
$$p(z|x) = \frac{p(x|z) \cdot p(z)}{p(x)}$$

posterior. evidence.



Generative Model.

($\sim 0.000 \cdots 1\%$)



So, we optimize $p(x|z)$

the likelihood.

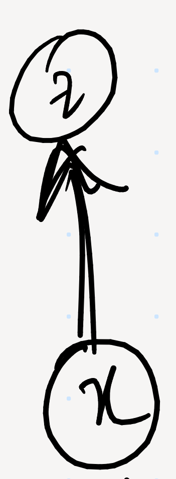
prior dist $p_\theta(z)$.

Setup :-

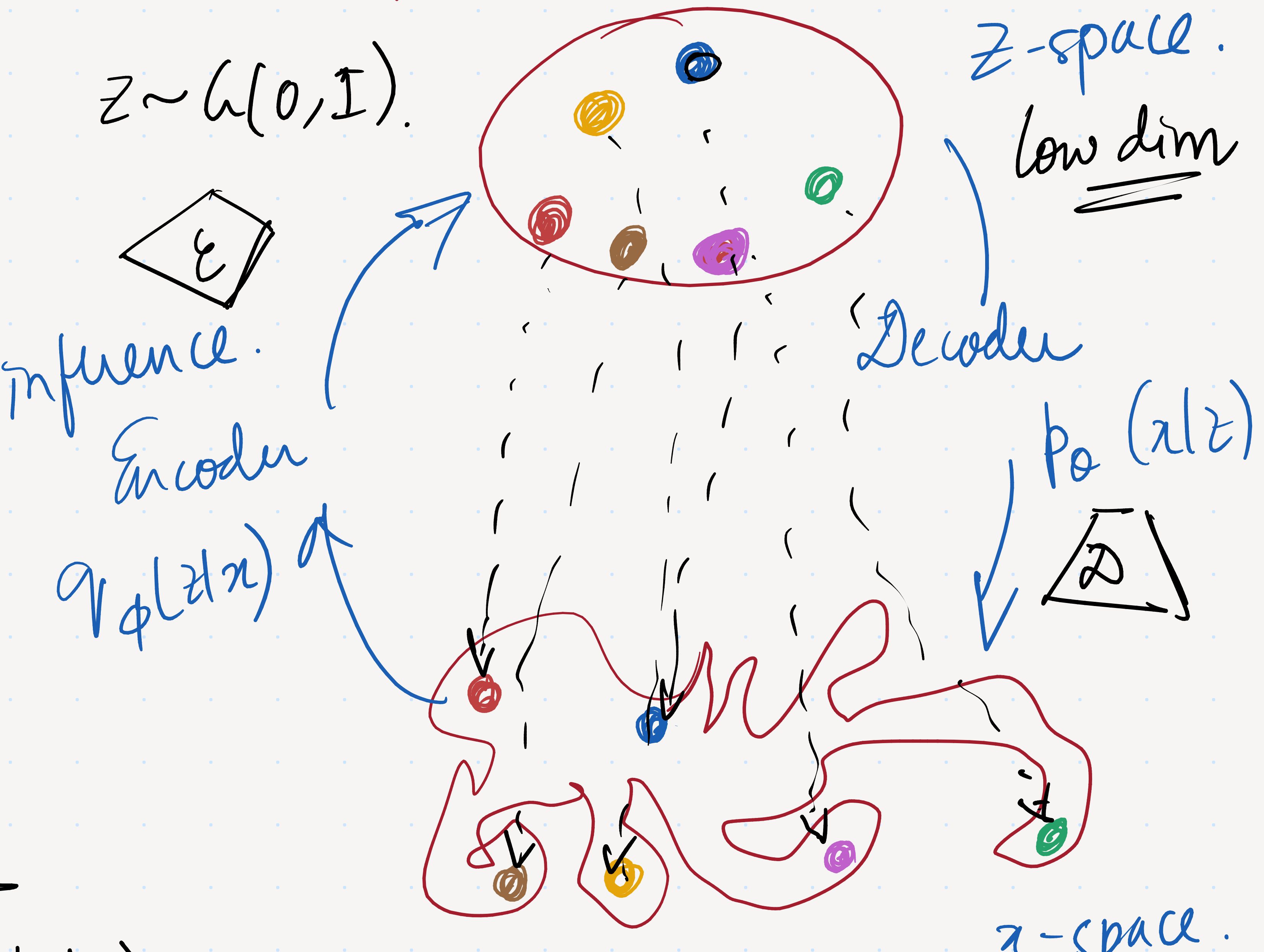
$z \sim \mathcal{N}(0, I)$.

$z \sim p(z) \to$ prior.

$x \sim p_\theta(x|z) \to$ Decoder.



Use $\boxed{q_\phi(z|x)}$ as a proxy for the encoder.



intractable to compute $p_\theta(z|x) \to$ intractable $\sim q_\phi(z|x)$.

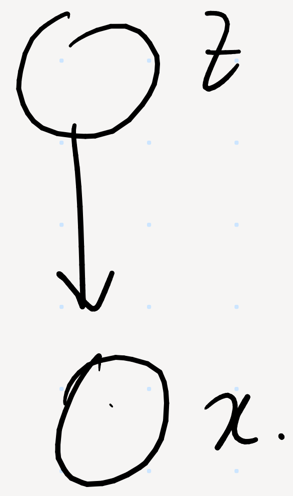
high Dim

Dataset $\mathcal{D}$.

We will find the lower bound to $p_\theta(x)$

$$p_\theta(z|x) = \frac{p_\theta(z,x)}{p_\theta(x)}$$

$$\log p_\theta(x) = \mathbb{E}_{z\sim q_\phi(z|x)} \log p_\theta(x).$$



$$= \mathbb{E}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(z,x)}{p_\theta(z|x)}$$

$$= \mathbb{E}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(z,x)}{q_\phi(z|x)} \cdot \frac{q_\phi(z|x)}{p_\theta(z|x)}$$

$$= \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)} \log \frac{\overset{\text{intractable}}{p_\theta(z,x)}}{q_\phi(z|x)}}_{L_{\theta,\phi}(x)} + \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)} \log \frac{q_\phi(z|x)}{p_\theta(z|x)}}_{D_{KL}\left(q_\phi(z|x) \,\|\, \underset{\text{intractable.}}{p_\theta(z|x)}\right)}$$

VLB, ELBO

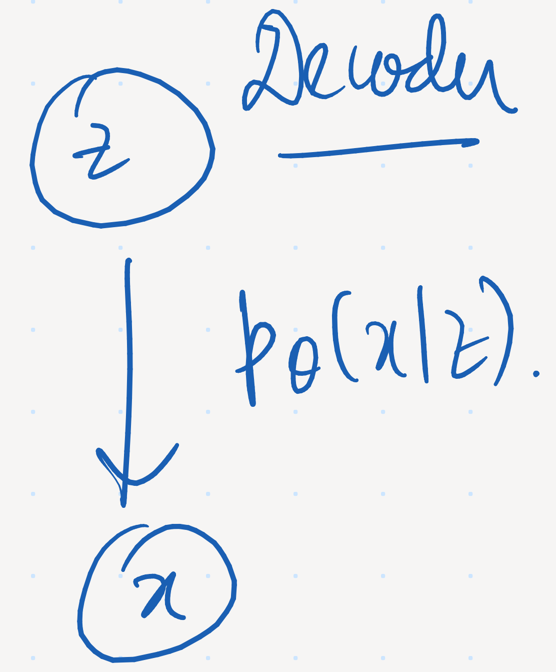
Variational Lower Bound

Evidence Lower Bound.

We will optimize this instead

$$\mathbb{E}_{z \sim q_\phi(z|x)} \log \frac{p_\theta(z,x)}{q_\phi(z|x)}$$

$p_\theta(z,x) \to$ intractable.



$\mathcal{L}_{\theta,\phi}(x)$

tractable

Decoder

simple

$$\Rightarrow \mathbb{E}_{z \sim q_\phi(z|x)} \log \frac{p_\theta(x|z) \cdot p_\theta(z)}{q_\phi(z|x)}$$

$p_\theta(z) \sim$ Normal.

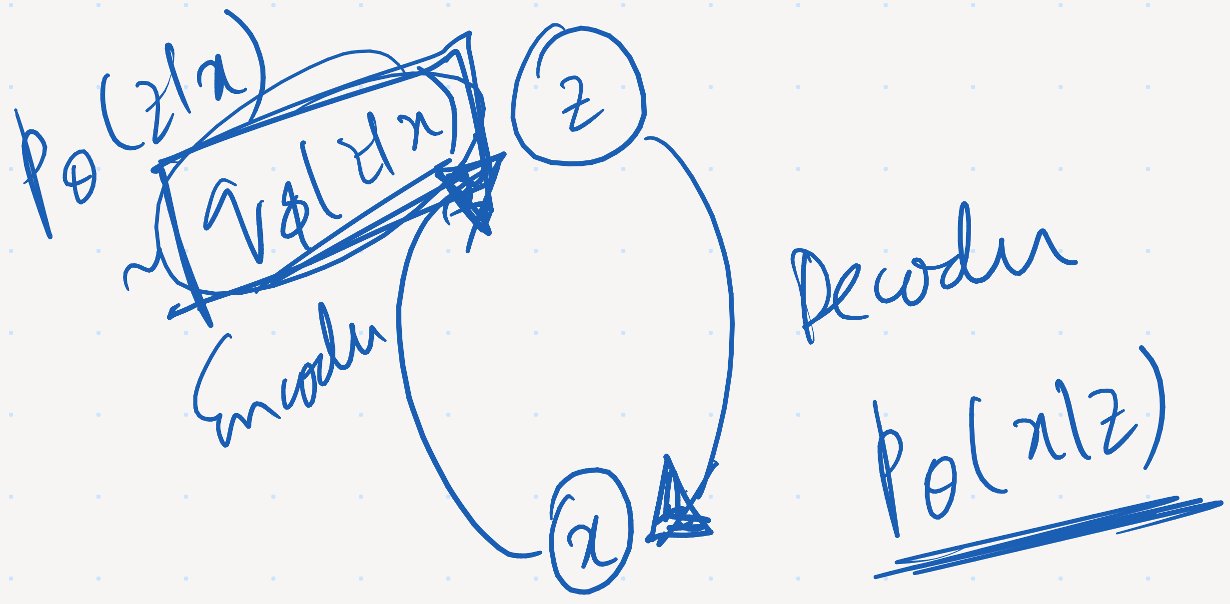
surrogate.

$$p_\theta(x|z) = \frac{p_\theta(z,x)}{p_\theta(z)}$$

$$p_\theta(z,x) = p_\theta(x|z) \cdot p_\theta(z).$$

$$\Rightarrow \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z)}_{\text{(Reconstruction error)}} + \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \log \frac{p_\theta(z)}{q_\phi(z|x)}}_{-D_{KL}(q_\phi(z|x) \,\|\, p_\theta(z))}$$

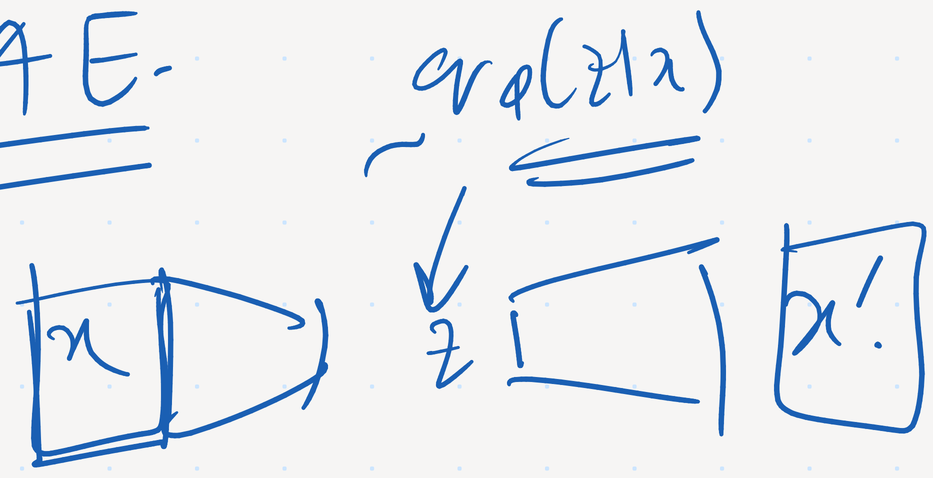
Regularization



$x \to$ Dataset.

$x \to$ Generate via(z).

AE.



$$\min \|x - x'\|_2$$

$$\begin{tikzcd} z \arrow[d] \\ x \end{tikzcd}$$

$$p_\theta(x|z) = \mathcal{N}(x; \underbrace{G_\theta(z)}_{\mu}, \underbrace{\eta}_{\sigma} I)$$ → As assumed before.

$$\log p_\theta(x|z) = ?$$

$$G(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

$$\Rightarrow \underbrace{\cancel{\log\left(\frac{1}{\sqrt{2\pi}}\right)}}_{\text{Constant}} + \underbrace{\cancel{\log\left(\frac{1}{\sqrt{\sigma^2}}\right)}}_{\text{Constant}} + \log_e e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow -\frac{1}{2\sigma^2}\|x-\mu\|^2$$

$$\log p_\theta(x|z) \Rightarrow -\frac{1}{2\eta^2}\|x - G_\theta(z)\|^2 + \text{constant.}$$

expected L_2-reconstruction error of the encoder model.

Maximizing the VLB encourages q_ϕ to be point mass.

$$D_{KL}(q\|p) = \mathbb{E}_{z\sim q} \log \frac{q(z)}{p(z)}$$

- $D_{KL}(q\|p) \neq D_{KL}(p\|q)$
- Measure of how far p is from q, hence if $p \sim q$ then $D_{KL}(p\|q) = 0$
- $D_{KL}(q\|p) \geq 0$ & is $= 0$ if $p = q$.

$$\mathcal{L}_{\phi,\theta}(x) = \mathbb{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) + \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(z)}{q_\phi(z|x)}}_{-D_{KL}(q_\phi(z|x)\|p_\theta(z))}$$

max $\mathcal{L}$ → min D_{KL}

hence $D_{KL} = 0$ when $\underline{\underline{p = q}}$

So maximizing the VLB, $\mathcal{L}_{\phi,\theta}$ pushes $q_\phi(z|x)$ towards $p(z)$, prevents q_ϕ from being a point $\underline{\underline{\text{mass}}}$.

$$\log p_\theta(x) = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x)\|p(z))}_{\mathcal{L}_{\theta,\phi}(x)} + \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)} \log \frac{q_\phi(z|x)}{p_\theta(z|x)}}_{D_{KL}(q_\phi(z|x)\|p_\theta(z|x))}$$

So, Maximizing VLB, $L_{\theta,\phi}$:-

- Maximizes $p(x) \rightarrow$ roughly
- Minimizes KL divergence b/w $q_\phi(z|x)$ & $p_\theta(z|x)$ making q_ϕ better.

Instead of optimizing $\sum_{i=1}^{n} \log p_\theta(x_i)$, we optimize.

$$\sum_{i=1}^{n} L_{\phi,\theta}(x_i)$$

We optimize the lower bound to be max.

$$\text{w/} \quad L_{\theta,\phi}(x) = \underbrace{\mathbb{E}_{z \sim q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)}}_{\text{Lower bound.}}$$

← intractable.

Optimizing VLB

$$\max_{\theta,\phi} \sum_{i=1}^{n} L_{\theta,\phi}(x_i).$$

$q_\phi(z|x) \sim p_\theta(z|x)$ $\qquad$ $p_\theta(x|z)$.

One possibility → for each x_i find the best $q_\phi(z|x_i)$ by multiple gradient steps in ϕ. Then gradient ascent in θ.

This results in expensive inference steps. / updates.

Instead

Amortize the inference costs by learning an inference n/w.

n/w

$x \to (\mu, \Sigma)$ w/ $q_\phi(z|x) = \mathcal{N}(z; \mu(x), \Sigma(x))$

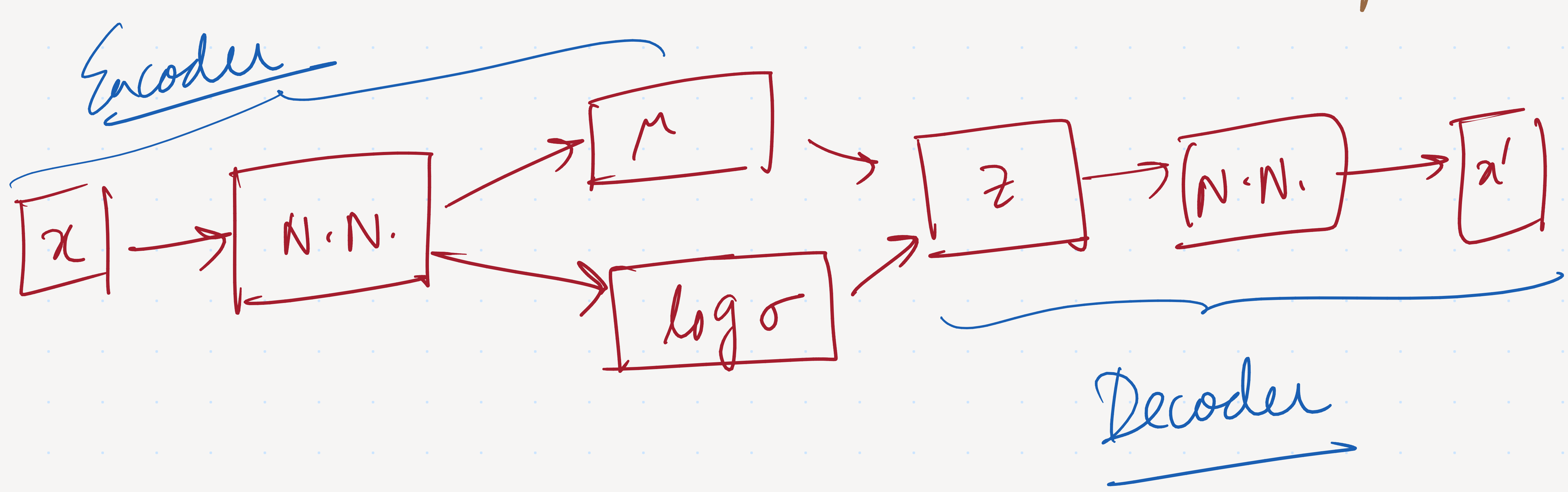
or $\sigma(x)I$

Parameters of the inference models are shared b/w the data-points

Reparameterization Trick.

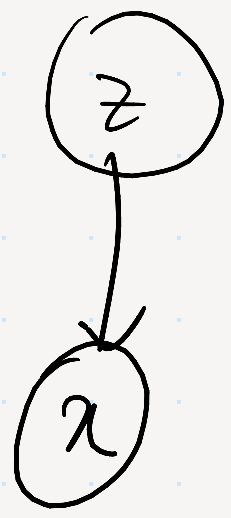
Separate random sources from differentiable quantities.

VAE-Architecture



Stochastic Gradient Optimization of the VLB.

Dataset $\mathcal{D} = \{x_i\}_{i=1,2,3,\ldots n}$



Solve,

$$\max_{\theta,\phi} \sum_{x_i \in \mathcal{D}} \mathcal{L}_{\theta,\phi}(x_i)$$

intractable.

w/ $$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{z \sim q_\phi(z|x)} \left(\log \frac{p_\theta(x,z)}{q_\phi(z|x)} \right)$$

Computing $\nabla_{\theta,\phi} \mathcal{L}_{\theta,\phi}(x_i)$ is intractable, but there are unbiased estimators. We need to survey all of z that are possible & then differentiate b/w θ & ϕ, and hence this is intractable.

Easy to get unbiased $\nabla_\theta \mathcal{L}_{\theta,\phi}$:-

$$\nabla_\theta \mathcal{L}_{\theta,\phi}(x) = \nabla_\theta \mathbb{E}_{z \sim q_\phi(z|x)} [\log p_\theta(x,z) - \log q_\phi(z|x)]$$

$$= \mathbb{E}_{z \sim q_\phi(z|x)} \nabla_\theta [\log p_\theta(x,z) - \log q_\phi(z|x)]$$

$$= \mathbb{E}_{z \sim q_\phi(z|x)} \nabla_\theta \log p_\theta(x, z).$$

$$\approx \underbrace{\nabla_\theta \log p_\theta(x, z)}_{\text{unbiased estimate.}} \rightarrow w/ \ z \sim q_\phi(z|x)$$

Sample from $\mathcal{D}$ and evaluate at randomly chosen z.

But it is not easy to get an unbiased estimate of $\nabla_\phi \mathcal{L}_{\theta,\phi}$ since ∇ doesn't commute.

$$\nabla_\phi \mathcal{L}_{\theta,\phi}(x) = \nabla_\phi \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log p_\theta(x, z) - \log q_\phi(z|x) \right]$$

$$\neq \mathbb{E}_{z \sim q_\phi(z|x)} \nabla_\phi \left(\log p_\theta(x, z) - \log q_\phi(z|x) \right)$$

Since, $q_\phi(z|x)$ will keep on changing, when we differentiate w.r.t. ϕ

Recall, $q_\phi(z|x) = \mathcal{N}(z; \mu(x), \sigma(x) I)$

$$= \mu(x) + \sigma(x) \cdot \varepsilon, \ w/ \ \varepsilon \sim \mathcal{N}(0, I).$$

$$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{z\sim q_\phi(z|x)}\left(\log p_\theta(x,z) - \log q_\phi(z|x)\right)$$

$$= \mathbb{E}_{\varepsilon\sim p(\varepsilon)}\left(\log p_\theta(x,z) - \log q_\phi(z|x)\right)$$

Form an estimator of $\mathcal{L}_{\theta,\phi}(x)$ as $\tilde{\mathcal{L}}_{\theta,\phi}(x)$ by :-

$\varepsilon\sim p(\varepsilon)$

$$z = \mu_\phi(x) + \sigma_\phi(x)\cdot\varepsilon = g(\phi, x, \varepsilon)$$

$$\hat{\mathcal{L}}_{\theta,\phi}(x) = \log p_\theta(x,z) - \log q_\phi(z|x)$$

Unbiased estimate of $\nabla_\phi \mathcal{L}_{\theta,\phi}(x)$:-

$$\nabla_\phi \hat{\mathcal{L}}_{\theta,\phi}(x)$$

Note :- $\mathbb{E}_{\varepsilon\sim p(\varepsilon)} \hat{\mathcal{L}}_{\theta,\phi}(x) = \mathcal{L}_{\theta,\phi}(x)$

So, $\mathbb{E}_{\varepsilon\sim p(\varepsilon)} \nabla_\phi \tilde{\mathcal{L}}_{\theta,\phi}(x) = \nabla_\phi \mathbb{E}_{\varepsilon\sim p(\varepsilon)} \hat{\mathcal{L}}_{\theta,\phi} = \nabla_\phi \mathcal{L}_{\theta,\phi}$

optimize VAE params. w.r.t. stochastic Gradient.