Diffusion Models

Caetano Müller - GMAP

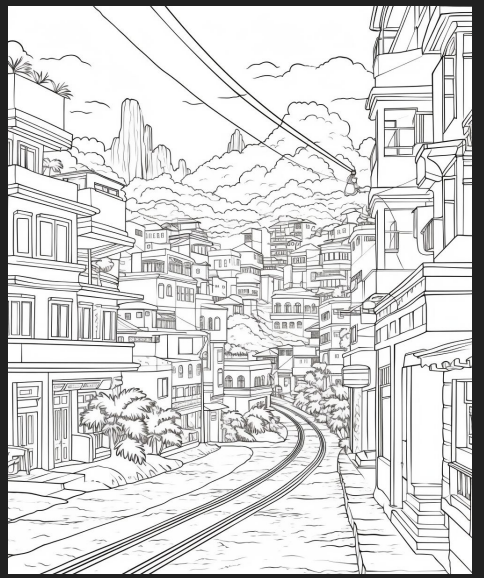
Introduction

What are diffusion models?

Generative Models?

How do they work?

a beautiful, colorful table spread of smoked meats



Coloring page for kids, street scene of Rio de Janeiro Brazil, cartoon style, thick lines, low detail, no shading --ar 9:11

Introduction

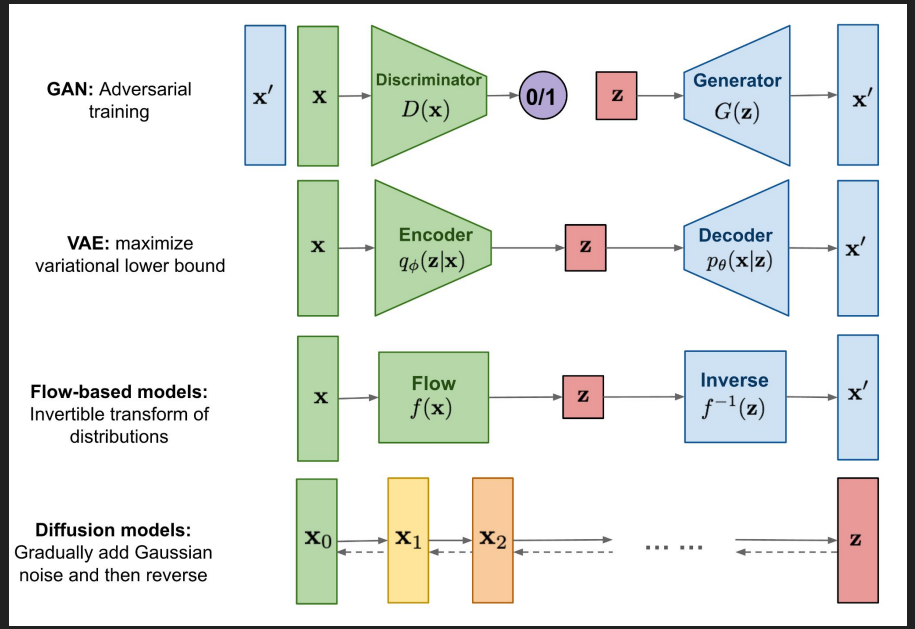
Why are these models relevant to us?

Introduction

- KL-Divergence
- Variational Inference
- Autoencoders and Variational Autoencoders
- Generative Adversarial Networks
- Diffusion Models

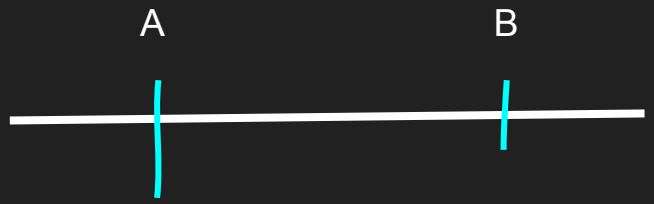
Introduction

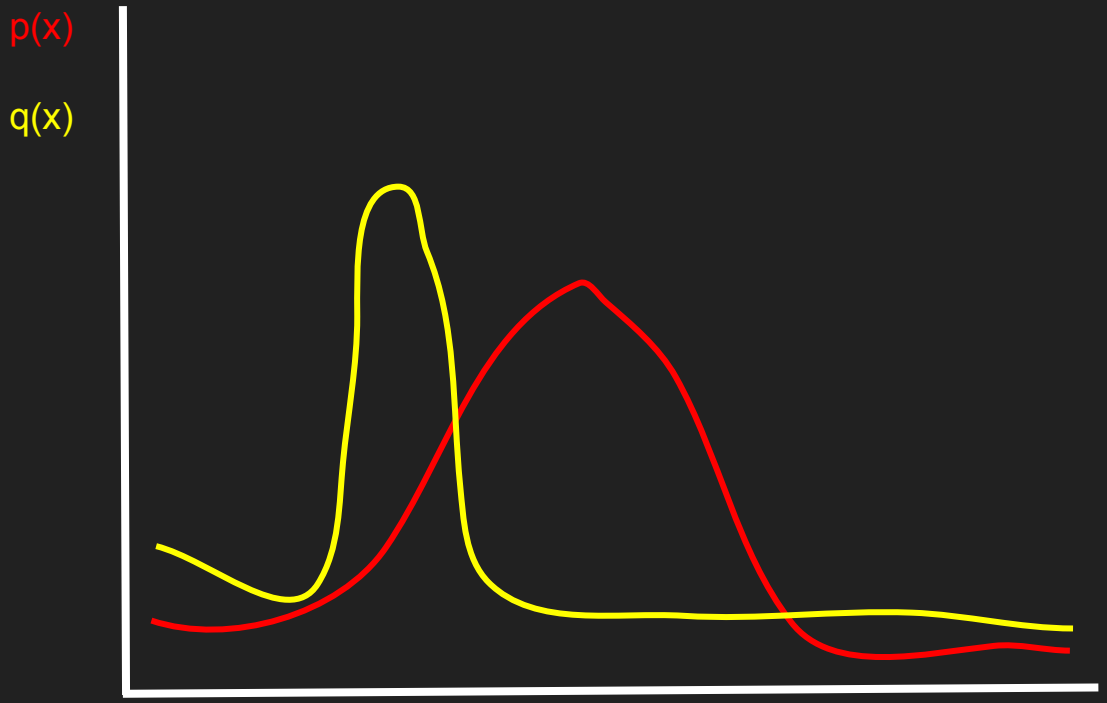
- Coding diffusion model in python
- Try it yourself on Colab
- :)



Kullback-leibler Divergence (KL-Divergence)

Represents the divergence between two distributions





KL - Divergence

Discrete

$$D_{\mathrm{KL}}(p(x)||q(x)) = \sum_{x \in X} p(x) ln \quad \frac{p(x)}{q(x)}$$

Continuous

$$D_{\mathrm{KL}}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) ln \, \frac{p(x)}{q(x)} \, dx$$

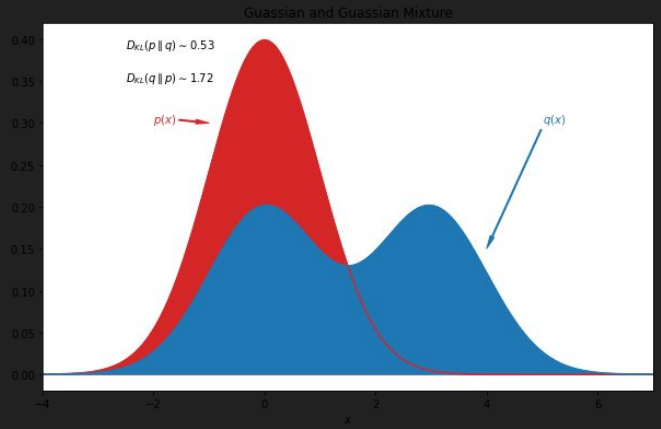
KL-Divergence

The order of the parameters change the equation therefore it is not precisely a distance between two distributions since the $D_{kl}(p(x) \,||\, q(x)) \neq D_{kl}(q(x) \,||\, p(x))$

$$D_{\mathrm{KL}}(p(x)\,||\,q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Expectation

$$D_{\mathrm{KL}}(p(x)\,||\,q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)}\, dx$$



KL-Divergence - Example

Unfair Coin Distribution $C \in \{H, T\}$ can be either Heads (H) or Tails (T),

therefore C can be distributed in a Bernoulli Distribution.

$C \sim B(\Theta)$ so that Θ is the probability of H.

$p(x) = B(0.6)$ $q(x) = B(0.7)$

KL-Divergence - Example

Unfair Coin Distribution C ∈ {H, T} can be either Heads (H) or Tails (T),

therefore C can be distributed in a Bernoulli Distribution.

C ~ B(Θ) so that Θ is the probability of H.

p(x) = B(0.6) q(x) = B(0.7)

$$D_{\mathrm{KL}} = \sum_{c=0}^{1} p(c) \ln \left(\frac{p(c)}{q(c)} \right)$$

KL-Divergence - Example

Unfair Coin Distribution C ∈ {H, T} can be either Heads (H) or Tails (T),

therefore C can be distributed in a Bernoulli Distribution.

C ~ B(Θ) so that Θ is the probability of H.

p(x) = B(0.6) q(x) = B(0.7)

$$D_{\mathrm{KL}} = \sum_{c=0}^{1} p(c) \ln\left(\frac{p(c)}{q(c)}\right)$$

$$D_{\mathrm{KL}} = (1 - 0.6) \ln\left(\frac{1-0.6}{1-0.7}\right) + 0.6 \ln\left(\frac{0.6}{0.7}\right)$$

KL-Divergence - Example

Unfair Coin Distribution C ∈ {H, T} can be either Heads (H) or Tails (T),

therefore C can be distributed in a Bernoulli Distribution.

C ~ B(Θ) so that Θ is the probability of H.

p(x) = B(0.6) q(x) = B(0.7)

$$D_{\mathrm{KL}} = \sum_{c=0}^{1} p(c) \ln\left(\frac{p(c)}{q(c)}\right)$$

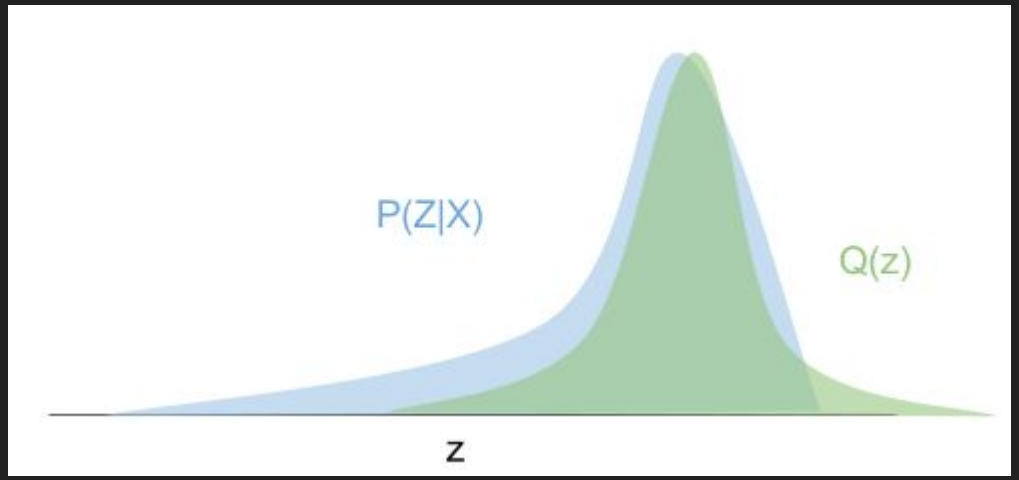
$$D_{\mathrm{KL}} = (1 - 0.6) \ln\left(\frac{1-0.6}{1-0.7}\right) + 0.6 \ln\left(\frac{0.6}{0.7}\right)$$

$$D_{\mathrm{KL}} = 0.0225824$$

Both Kullback and Leibler were cryptanalysts

Variational Inference

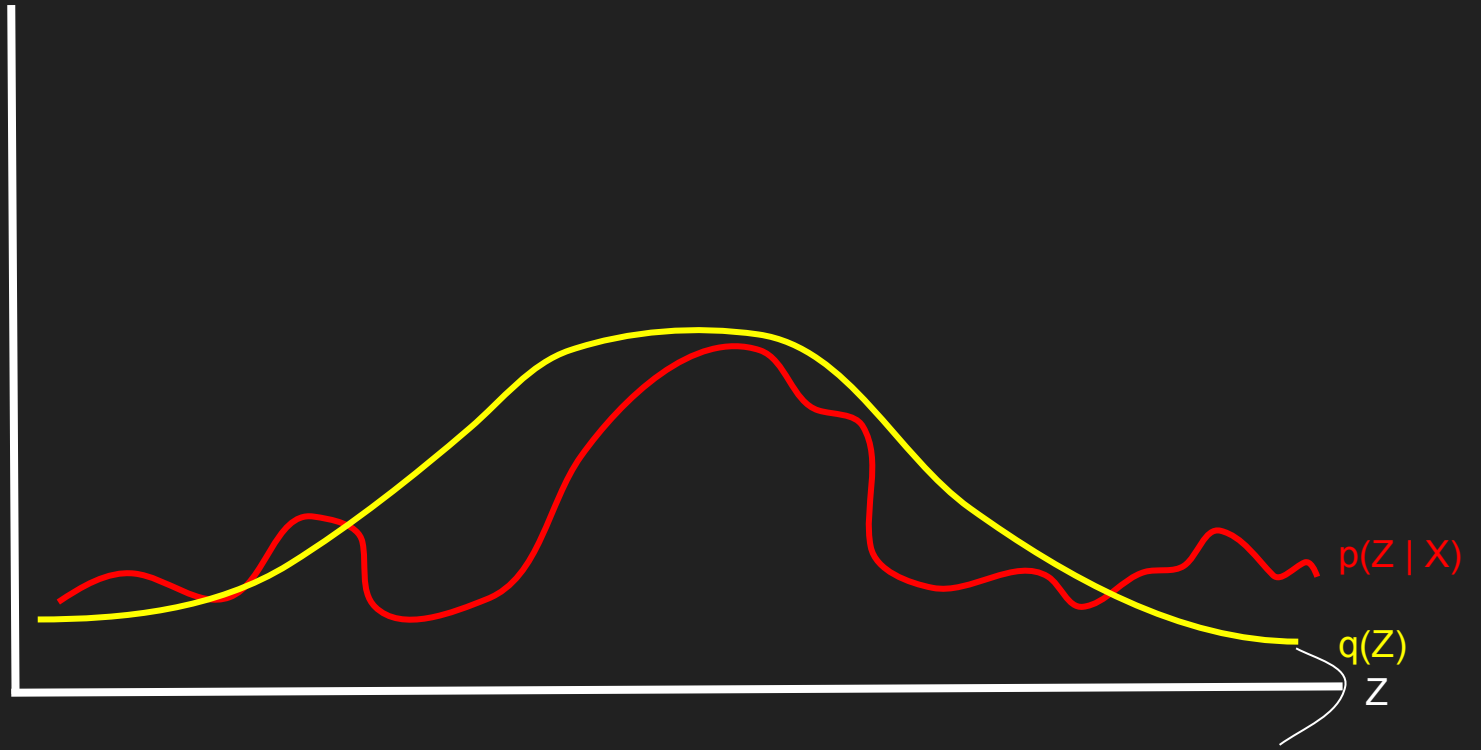
The Idea behind variational Inference is to be able to work with some distribution of data that we are not capable of deriving and extracting valuable information from directly



Variational Inference

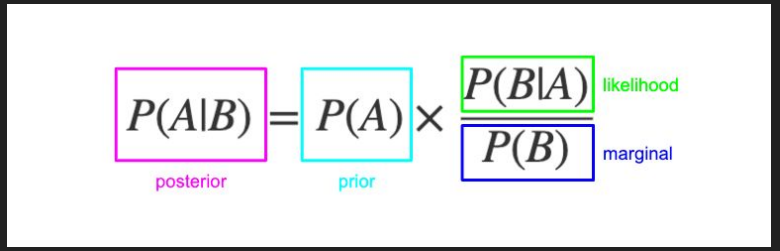
Z is our latent Variable and we would like to infer it (for our VAE Images)

We want to create this surrogate and get it as close as we can to the actual distribution

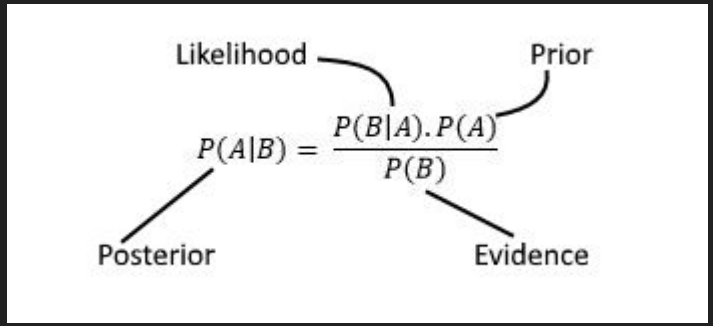


Variational Inference

Evidence is Intractable so we must find a way Around it



What we really want is $q(Z) \approx p(Z \mid X = D)$



Variational Inference

We will use de KL-Divergence to calculate and minimize our Loss

So what we actually want is the best surrogate for us:

$q^*(Z) = argmin(D_{kl} (q(Z) || p(Z | X = D)))$

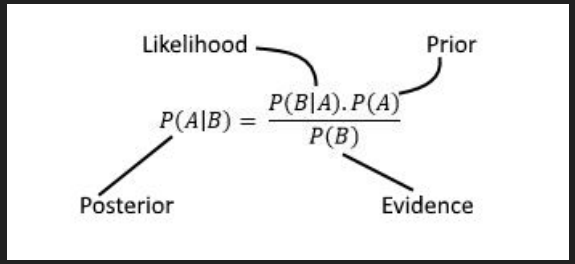
$E = E_{Z \sim q(Z)}$ which is the expectation of Z distributed over q(Z)

$$D_{\mathrm{KL}}(q(Z)||p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

But again we have a problem since we are need the posterior in *p(Z | D)*

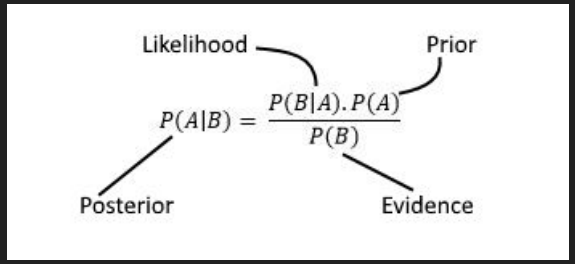
$\mathbb{E}[\ln(q(Z|X))]$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$



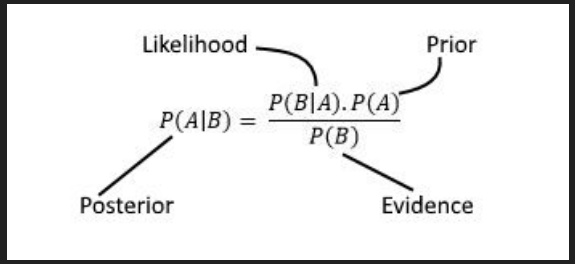
$$\mathbb{E}[\ln(q(Z|X))]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$



$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln(q(Z))\right] - \mathbb{E}\left[\ln(p(X,Z))\right] + \mathbb{E}\left[\ln(p(X))\right]$$

$$\mathbb{E}[\ln(q(Z|X))]$$

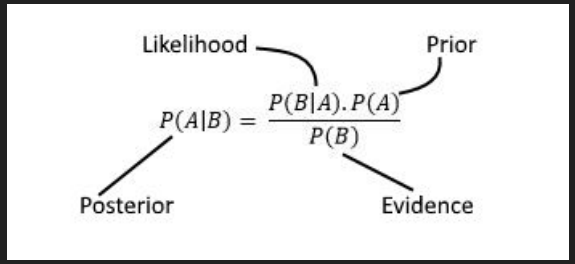


$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln(q(Z))\right] - \mathbb{E}\left[\ln(p(X,Z))\right] + \mathbb{E}\left[\ln(p(X))\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$\mathbb{E}[\ln(q(Z|X))]$



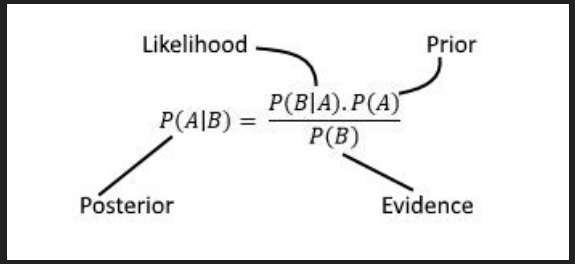
$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln(q(Z))\right] - \mathbb{E}\left[\ln(p(X,Z))\right] + \mathbb{E}\left[\ln(p(X))\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$\mathbb{E}[\ln(q(Z|X))]$



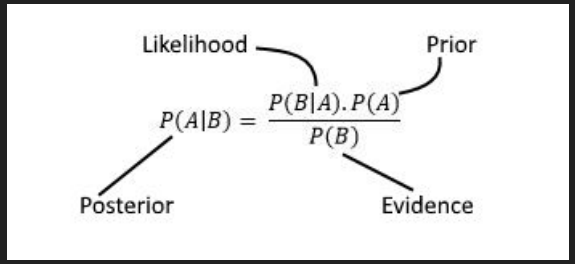
$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln(q(Z))\right] - \mathbb{E}\left[\ln(p(X,Z))\right] + \mathbb{E}\left[\ln(p(X))\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$\mathbb{E}[\ln(q(Z|X))]$



$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

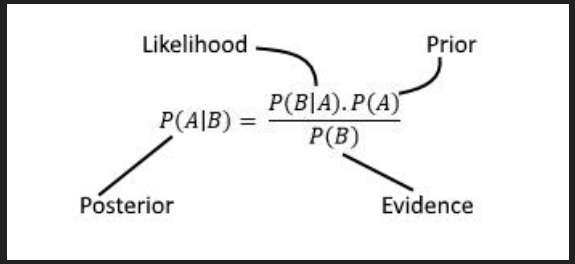
$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln(q(Z))\right] - \mathbb{E}\left[\ln(p(X,Z))\right] + \mathbb{E}\left[\ln(p(X))\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X))$$

$\mathbb{E}[\ln(q(Z|X))]$



$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln(q(Z))\right] - \mathbb{E}\left[\ln(p(X,Z))\right] + \mathbb{E}\left[\ln(p(X))\right]$$

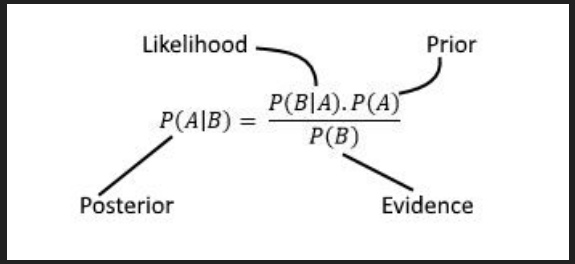
$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X))$$

$$\ln(p(X)) \geq \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + D_{\mathrm{KL}}(q(Z)\|p(Z|X=D))$$

$\mathbb{E}[\ln(q(Z|X))]$



$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln(q(Z))\right] - \mathbb{E}\left[\ln(p(X,Z))\right] + \mathbb{E}\left[\ln(p(X))\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

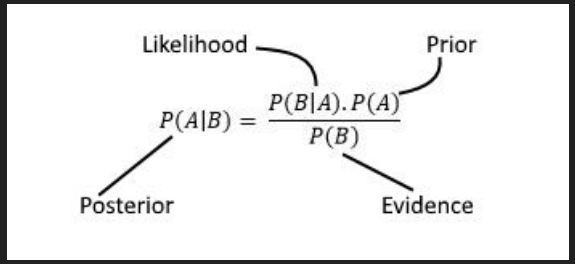
$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X))$$

$$\ln(p(X)) \geq \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + D_{\mathrm{KL}}(q(Z)\|p(Z|X=D))$$

ELBO

KL-Divergence must be > 0

$\mathbb{E}[\ln(q(Z|X))]$



$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln(q(Z))\right] - \mathbb{E}\left[\ln(p(X,Z))\right] + \mathbb{E}\left[\ln(p(X))\right]$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X)) \int q(Z|X,Z)\,\mathrm{d}Z$$

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + \ln(p(X))$$

$\mathbb{E}[\ln(q(Z|X))]$

$$\ln(p(X)) \geq \mathbb{E}[\ln(q(Z))] - \mathbb{E}[\ln(p(X,Z))] + D_{\mathrm{KL}}(q(Z)\|p(Z|X=D))$$

ELBO

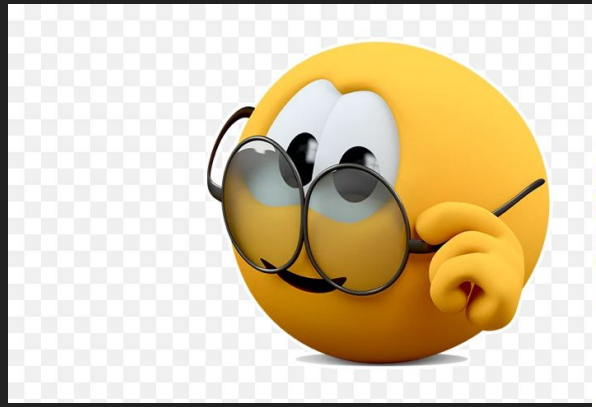
KL-Divergence must be > 0

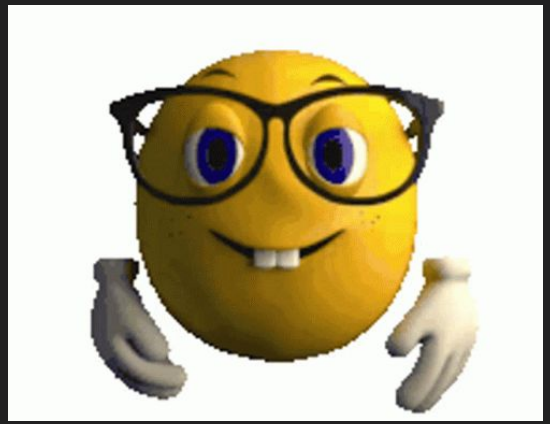
Variational INference - ELBO

$$ELBO = L(q) = \mathbb{E}\left[\ln\left(\frac{p(Z,X)}{q(Z)}\right)\right]$$

q(Z) = argmin(D_{kl} (q(Z) || p(Z | X = D))) ≡ argmax(L(q(Z)))

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{L}(q(Z) + \ln(p(Z))$$

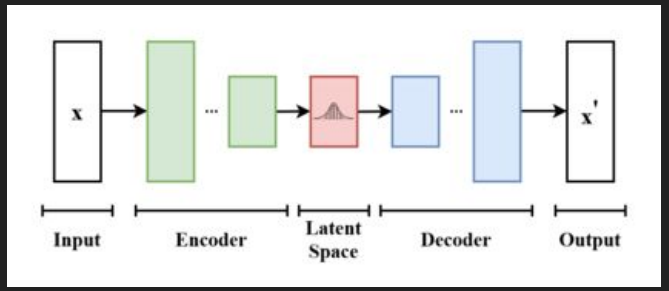


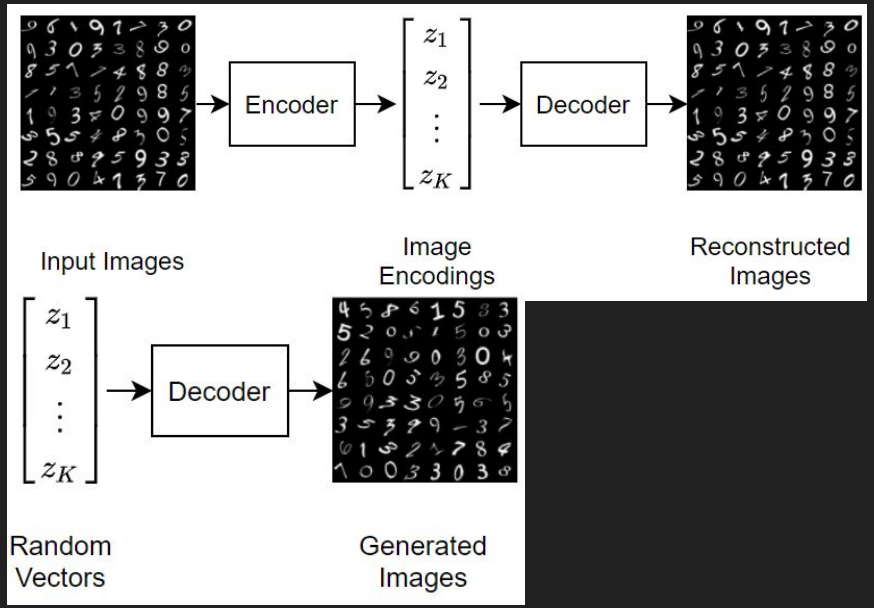


Variational Autoencoder (VAE)

Basically combines the fields of Autoencoders (NNs) and Variational Inference (SGA)

We use it to basically generate new data from its latent space (mainly)

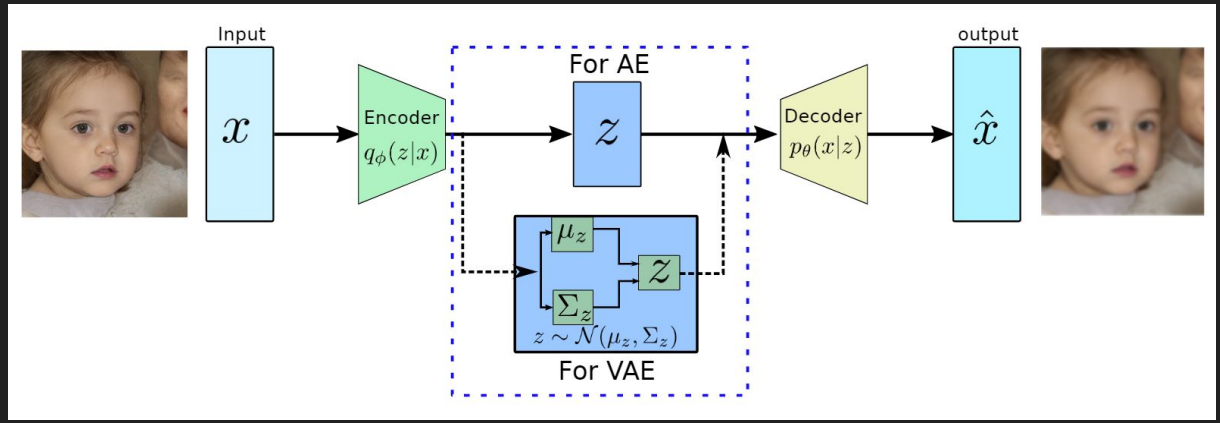




VAE

What is an autoencoder (AE)?

AEs are mainly used from compressing data, de input data goes through the encoder and then we can just decompress (decode) it (with some loss). This can be really used for dimensionality reduction.



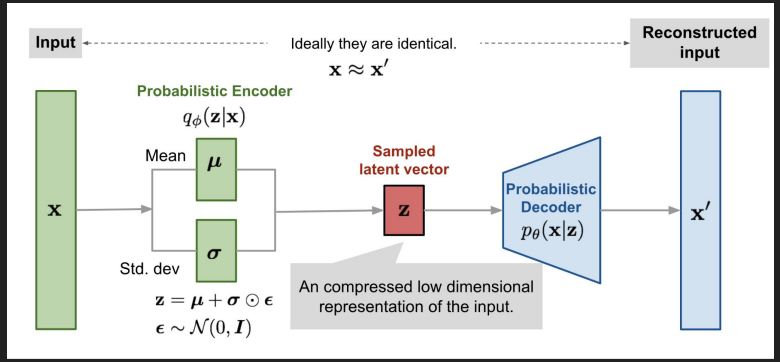
VAE

In the case of AEs the problem is how unregular and unstructured the data is stored in **Z**, therefore it is quite hard to generate new data going through the process of decoding.

w and w' are the weights of the encoder and decoder

$$F_w(X) = Z \quad \text{and} \quad G_{w'}(Z) = X$$

$$\Theta = ||X - G_{w'}(F_w(X))||^2$$



VAE

Reparameterization Trick.

We will use this trick to find the gradient with lower variance using a second distribution independent of the parameters of our actual distribution

$\epsilon \sim p'(\epsilon)$-> Neutral Distribution $Z = g(\epsilon, X)$-> Transformation

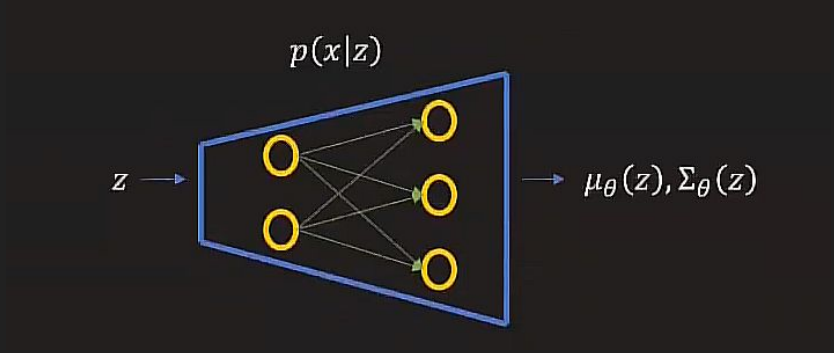
$$ELBO = L(q) = \mathbb{E}\left[\ln\left(\frac{p(Z,X)}{q(Z)}\right)\right]$$

$$\nabla \mathbb{E}_{p'}\left[\ln\left(p(X, g(\epsilon, X))\right) - \ln(q(g(\epsilon, X)))\right]$$

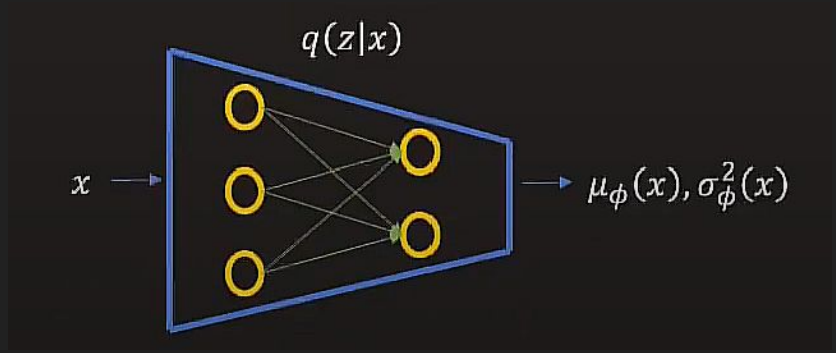
$$\nabla\left[\frac{1}{N}\sum_{i=1}^{N} \ln(p(X, g(\epsilon, X))) - \ln(q(g(\epsilon, X)))\right]$$

VAE

Updates likelihood



Updates Posterior



Now we just have to minimize our loss optimizing for both models above (Maximize ELBO)

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{L}(q(Z) + \ln(p(Z))$$

$$L = -\mathbb{E}_{q(Z)}\left[\ln\left(\frac{p(Z,X)}{q(Z)}\right)\right] + \ln(p(X))$$

Makes distribution similar to a Normal distribution (if using normal for approximation)

Similar to the regular AE and is trying to reconstruct data from latent space

https://lilianweng.github.io/posts/2018-08-12-vae/#reparameterization-trick


depositphotos
Image ID: 195433490
www.depositphotos.com

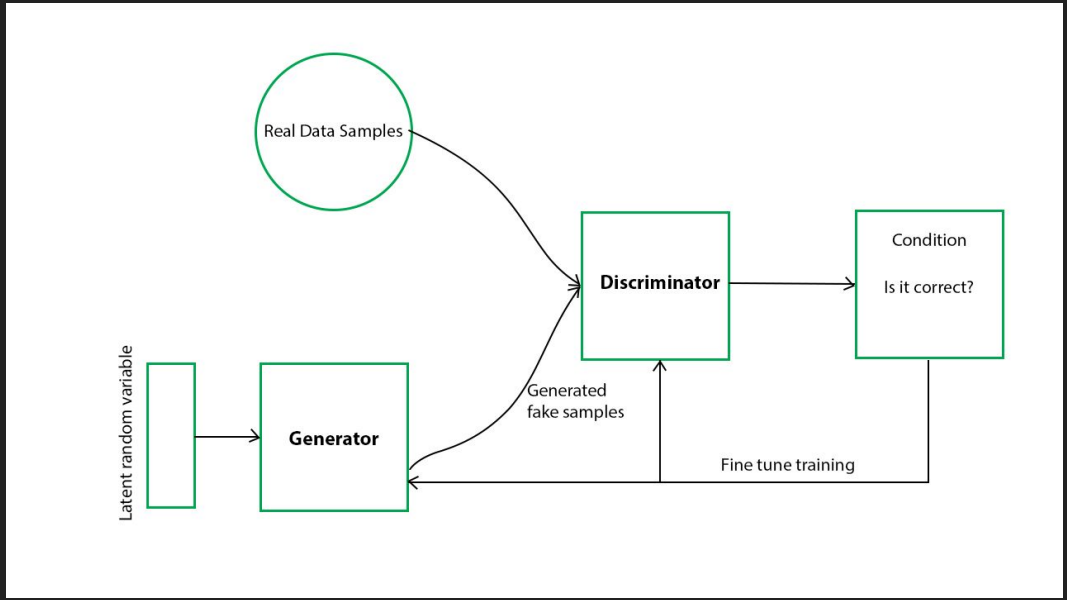

Nem gosto de mimosa
com narguet
ifunny.co

Generative Adversarial Networks (GANs)

The main Idea begin GANs is that they basically work like a game/competition between two models

G : Generator

D : Discriminator



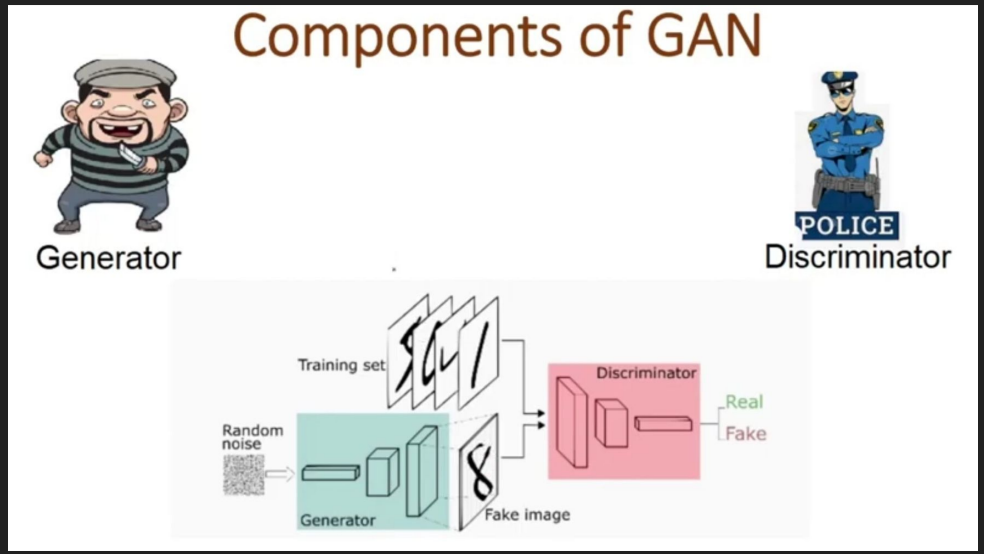
GANs

$$\min\max V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}}[\ln(D(x))] + \mathbb{E}_{z\sim p_z}[\ln(1-D(G(z)))]$$

$$\text{BCE} = -\sum_{i=1}^{N}\left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\right]$$

If y=1 then p = D(x)

If y=0 then p = D(G(z))



GANs

$$\operatorname{minimax} V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}}[\ln(D(x))] + \mathbb{E}_{z \sim p_z}[\ln(1 - D(G(z)))]$$

GANs

$$\min\max V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}}[\ln(D(x))] + \mathbb{E}_{z \sim p_z}[\ln(1 - D(G(z)))]$$

$$\text{BCE} = -\sum_{i=1}^{N} \left[y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i) \right]$$

GANs

$$\operatorname{minimax} V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}}[\ln(D(x))] + \mathbb{E}_{z \sim p_z}[\ln(1 - D(G(z)))]$$

$$\text{BCE} = -\sum_{i=1}^{N} \left[y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i) \right]$$

$$L = \ln(D(x)) + \ln(1 - D(G(z)))$$

GANs

$$\min\max V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}}[\ln(D(x))] + \mathbb{E}_{z \sim p_z}[\ln(1 - D(G(z)))]$$

$$\text{BCE} = -\sum_{i=1}^{N} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$$

$$L = \ln(D(x)) + \ln(1 - D(G(z)))$$

$$E[L] = \int p_{\text{data}}(x) \ln(D(x))\, dx + \int p_z(z) \ln(1 - D(G(z)))\, dz$$

GANs

$$\min\max V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}}[\ln(D(x))] + \mathbb{E}_{z \sim p_z}[\ln(1 - D(G(z)))]$$

$$\text{BCE} = -\sum_{i=1}^{N} \left[y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i) \right]$$

$$L = \ln(D(x)) + \ln(1 - D(G(z)))$$

$$E[L] = \int p_{\text{data}}(x) \ln(D(x))\, dx + \int p_z(z) \ln(1 - D(G(z)))\, dz$$

$$V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}}[\log(D(x))] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

GANs

Note that the Discriminator will try to maximize the loss function while the Generator will try to minimize it.

Ideally we want to maximize our Generator and given that with a fixed G the maximum for D(x) will be described by the expression $\frac{\mathrm{p}_{\mathrm{data}}(x)}{\mathrm{p}_{\mathrm{data}}(x)+\mathrm{p}_{\mathrm{g}}(x)}$

Now we can fix D(x) and try to find our minima of G

GANs

Given our known function

$$V(G,D) = \mathbb{E}_{x \sim p_{\text{data}}}[\log(D(x))] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

We can use our fixed D(x) = $\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x) + \mathrm{p}_{\text{g}}(x)}$

GANs

Given our known function

$$V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$$

We can use our fixed D(x) = $\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)}$

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(1 - \frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

GANs

Given our known function

$$V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$$

We can use our fixed D(x) = $\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)}$

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(1 - \frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

GANs

Given our known function

$$V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$$

We can use our fixed D(x) = $\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)}$

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(1 - \frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(\frac{\mathrm{p}_{\text{g}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

GANs - JS-Divergence

We will now use a new method to calculate the divergence between two distributions and that will be the Jensen-Shannon Divergence(JS-Divergence)

$$\mathrm{JS}(P||Q) = \frac{1}{2}\mathrm{DKL}(P||\frac{P+Q}{2}) + \frac{1}{2}\mathrm{DKL}(Q||\frac{P+Q}{2})$$

Now if we remember the D*kl* equation earlier displayed here

$$D_{\mathrm{KL}}(q(Z)\|p(Z|X=D)) = \mathbb{E}\left[\ln\left(\frac{q(Z)}{p(Z|D)}\right)\right]$$

And now changing those values we have the following

$$JS(P||Q) = \frac{1}{2}(\mathbb{E}_{x\sim P}[\log(\frac{2P}{P+Q})] + \mathbb{E}_{x\sim Q}[\log(\frac{2Q}{P+Q})])$$

GANs

Using the JS-Divergence to our known data of the Generator we have

$$V = \mathbb{E}_{x \sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x \sim p_g}[\log(\frac{\mathrm{p}_{\text{g}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

And seeing how close hour JS Divergence is to this formula

$$JS(P||Q) = \frac{1}{2}(\mathbb{E}_{x \sim P}[\log(\frac{2P}{P+Q})] + \mathbb{E}_{x \sim Q}[\log(\frac{2Q}{P+Q})])$$

With just a simple algebraic trick we can change our V to

$$V = \mathbb{E}_{x \sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x \sim p_g}[\log(\frac{\mathrm{p}_{\text{g}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

GANs

Using the JS-Divergence to our known data of the Generator we have

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(\frac{\mathrm{p}_{\text{g}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

And seeing how close hour JS Divergence is to this formula

$$JS(P||Q) = \frac{1}{2}(\mathbb{E}_{x\sim P}[\log(\frac{2P}{P+Q})] + \mathbb{E}_{x\sim Q}[\log(\frac{2Q}{P+Q})])$$

With just a simple algebraic trick we can change our V to

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{2\,\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(\frac{2\,\mathrm{p}_{\text{g}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

GANs

Using the JS-Divergence to our known data of the Generator we have

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(\frac{\mathrm{p}_{\text{g}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

And seeing how close hour JS Divergence is to this formula

$$JS(P||Q) = \frac{1}{2}(\mathbb{E}_{x\sim P}[\log(\frac{2P}{P+Q})] + \mathbb{E}_{x\sim Q}[\log(\frac{2Q}{P+Q})])$$

With just a simple algebraic trick we can change our V to

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(\frac{\mathrm{p}_{\text{g}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})]$$

$$V = \mathbb{E}_{x\sim p_{\text{data}}}[\log(\frac{2\mathrm{p}_{\text{data}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] + \mathbb{E}_{x\sim p_g}[\log(\frac{2\,\mathrm{p}_{\text{g}}(x)}{\mathrm{p}_{\text{data}}(x)+\mathrm{p}_{\text{g}}(x)})] \; -2log(2)$$

GANs

So finally we can Describe our V that minimizes G as

$$V = 2JS(p_{\text{data}}||p_g) - 2\log(2)$$

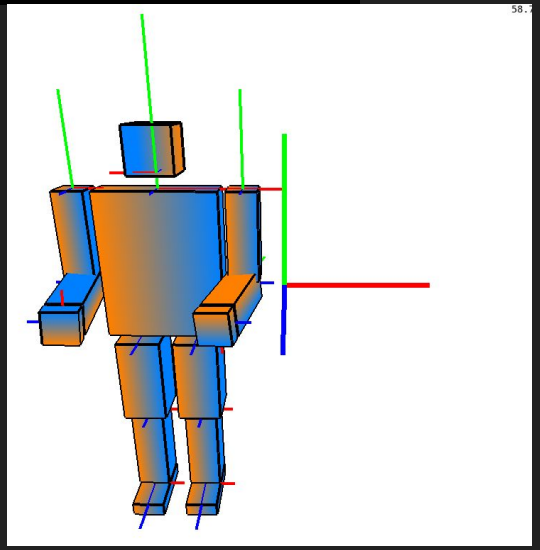
That mean that at our global minimum we have an ideal Generator with

$$p_{\text{data}} = p_g$$

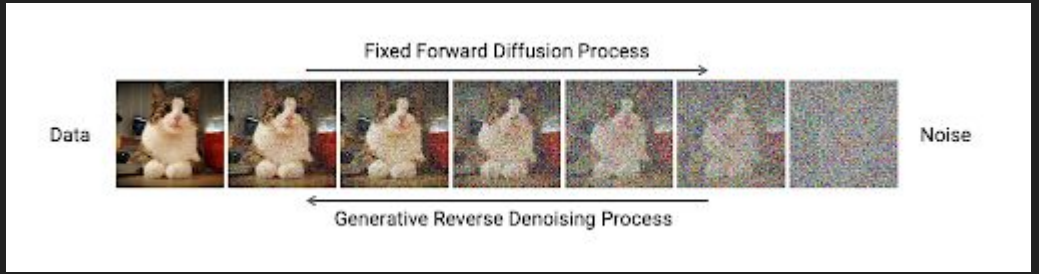
If this case ever occurs then we have our Discriminator as a useless classifier since it will be impossible to find the difference between real and generated data.

Note that training GANs is quite hard so this really is an ideal scenario


alamy
Image ID: 2AC1278
www.alamy.com



Diffusion Models



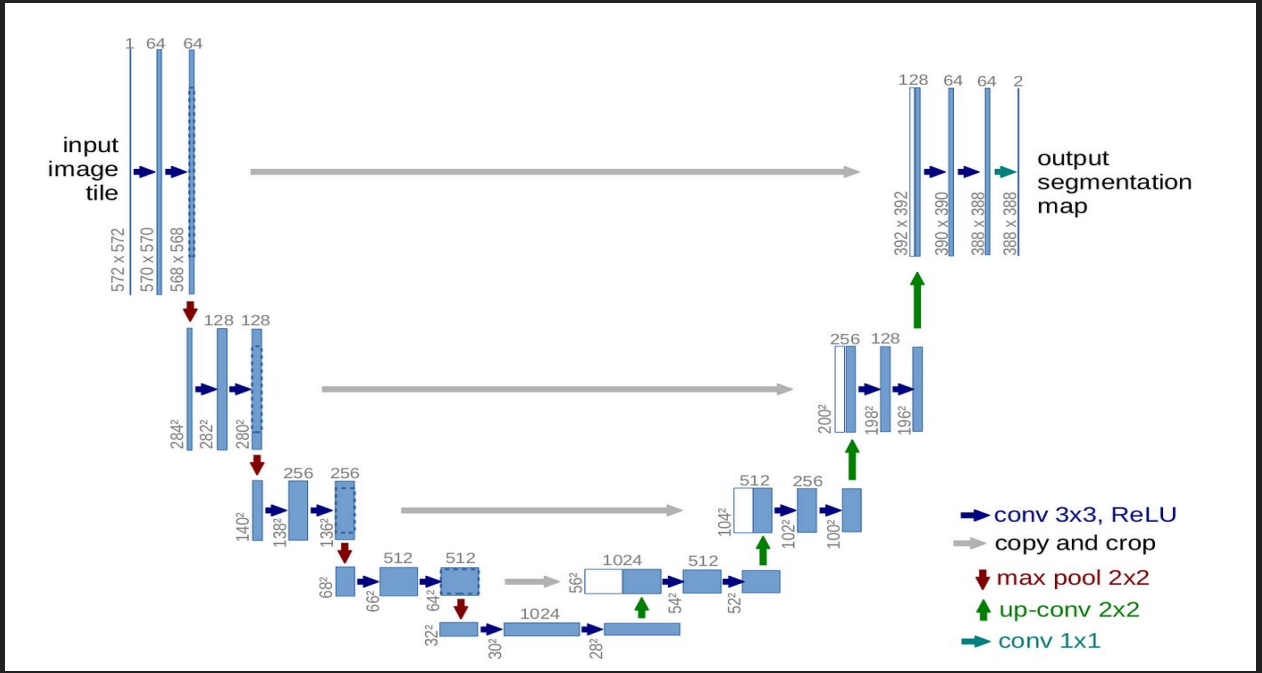
The Main Idea in Diffusion Models is to keep adding noise to an image following a Normal Distribution and then Generate on in these 3 options:

- Mean of Noise
- Original Image
- Noise of Image

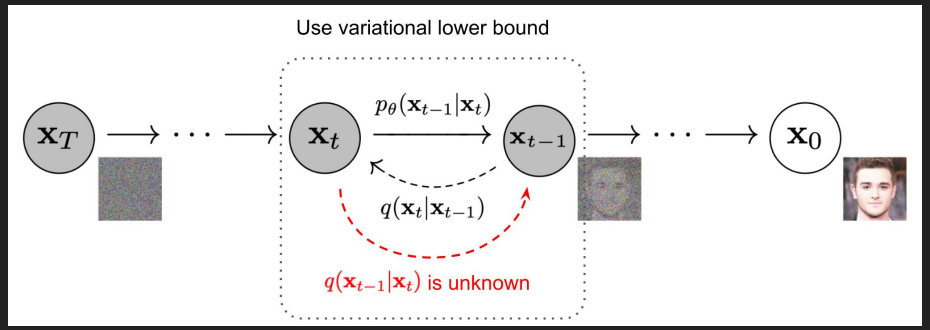
As we can already guess predicting the original image is not really the best option and later it will be shown that the 1st and 3rd option result in the same thing so we will just be predicting the noise directly

Diffusion Models

These models follow the U-NET architecture



Diffusion Models



Using the notation of

$x_t = Image$ And t represents time

Our forward process can be described by $q(x_t|x_{t-1})$

Whereas our reverse process can be described by $p(x_{t-1}|x_t)$

As said before *q* will follow a Normal Distribution so we can have it as

$$q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Beta scales between 0 and 1 and ensures the variance doesn't explode

Reparametrization Trick $= N(\mu, \sigma^2) = \mu + \sigma\epsilon$

Diffusion Models

$$q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

$$\alpha_t = 1 - \beta_t$$

We will again be using the reparametrization trick for the Normal $N(\mu, \sigma^2) = \mu + \sigma\epsilon$

$$\sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon$$

$$\bar{\alpha} = \prod_{i=1}^{t} \alpha_i$$

$$\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

Diffusion Models

Now we can describe the whole forward process of adding noise to the images with $\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$

Now we need to work on the reverse process of $p(x_{t-1}|x_t)$ with the notation

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t))$$

Using our loss function the negative log likelihood $-log(p_\theta(x_0))$

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

We are using the Evidence Lower Bound again because the log likelihood of x0 depends on all the x's that came before it and that is not tractable.

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}$$

$$\frac{p_\theta(x_0|x_{1:T})p_\theta(x_{1:T})}{p_\theta(x_0)}$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}$$

$$\frac{p_\theta(x_0|x_{1:T})p_\theta(x_{1:T})}{p_\theta(x_0)}$$

$$\frac{p_\theta(x_0,x_{1:T})}{p_\theta(x_0)}$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}$$

$$\log\left(\frac{q(x_{1:T}|x_0)}{\frac{p_\theta(x_0,x_{1:T})}{p_\theta(x_0)}}\right)$$

$$\frac{p_\theta(x_0|x_{1:T})p_\theta(x_{1:T})}{p_\theta(x_0)}$$

$$\frac{p_\theta(x_0,x_{1:T})}{p_\theta(x_0)}$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}$$

$$\log\left(\frac{q(x_{1:T}|x_0)}{\frac{p_\theta(x_0,x_{1:T})}{p_\theta(x_0)}}\right)$$

$$\frac{p_\theta(x_0|x_{1:T})p_\theta(x_{1:T})}{p_\theta(x_0)}$$

$$\log\left(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}\right) + \log(p_\theta(x_0))$$

$$\frac{p_\theta(x_0,x_{1:T})}{p_\theta(x_0)}$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0))$$

$$\log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + \log(p_\theta(x_0))$$

$$-log(p_\theta(x_0)) \leq -log(p_\theta(x_0)) + \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) + log(p_\theta(x_0))$$

$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$$

Variational Lower Bound

Diffusion Models

Now with the $-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$ equation we can work on minimizing it. We know that the term at the top is our forward process. As for the term at the bottom we can rewrite it as $p_\theta(x_0, x_{1:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$

So that we end up with the following:

Diffusion Models

Now with the $-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$ equation we can work on minimizing it. We know that the term at the top is our forward process. As for the term at the bottom we can rewrite it as $p_\theta(x_0, x_{1:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$

So that we end up with the following:

$$\log(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)})$$

Diffusion Models

Now with the $-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$ equation we can work on minimizing it. We know that the term at the top is our forward process. As for the term at the bottom we can rewrite it as $p_\theta(x_0, x_{1:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$

So that we end up with the following:

$$\log(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)})$$

Diffusion Models

Now with the $-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$ equation we can work on minimizing it. We know that the term at the top is our forward process. As for the term at the bottom we can rewrite it as $p_\theta(x_0, x_{1:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$

So that we end up with the following:

$$\log(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}) = -\log(p(x_T)) + \log(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)})$$

Diffusion Models

Now with the $-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$ equation we can work on minimizing it. We know that the term at the top is our forward process. As for the term at the bottom we can rewrite it as $p_\theta(x_0, x_{1:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$

So that we end up with the following:

$$\log(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}) = -\log(p(x_T)) + \log(\frac{\prod_{t=1}^{T} q(x_t|x_{t-1})}{\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)})$$

$$= -\log(p(x_T)) + \sum_{t=1}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)})$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})})$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) = -\log(p(x_T)) + \sum_{t=1}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)})$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) = -\log(p(x_T)) + \sum_{t=1}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)})$$

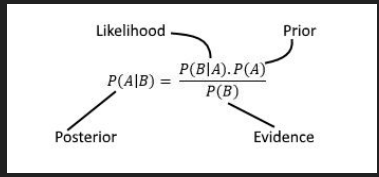
$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}) + \log(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)})$$

Diffusion Models

$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) = -\log(p(x_T)) + \sum_{t=1}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)})$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}) + \log(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)})$$

Diffusion Models

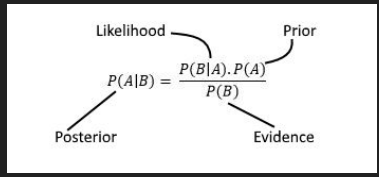


$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) = -\log(p(x_T)) + \sum_{t=1}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)})$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}) + \log(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)})$$

We now apply the Bayes and condition it on x0 to avoid high variances

Diffusion Models



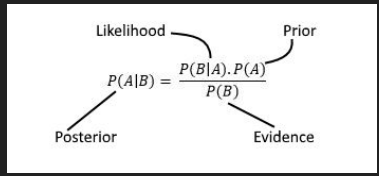
$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) = -\log(p(x_T)) + \sum_{t=1}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)})$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}) + \log(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)})$$

We now apply the Bayes and condition it on x0 to avoid high variances

$$q(x_t|x_{t-1}) = \frac{q(x_{t-1}|x_t)q(x_t)}{q(x_{t-1})}$$

Diffusion Models



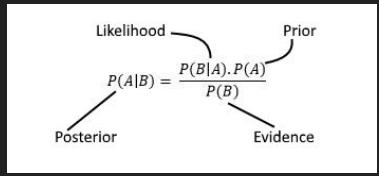
$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) = -\log(p(x_T)) + \sum_{t=1}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)})$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}) + \log(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)})$$

We now apply the Bayes and condition it on x0 to avoid high variances

$$q(x_t|x_{t-1}) = \frac{q(x_{t-1}|x_t)q(x_t)}{q(x_{t-1})} => \frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

Diffusion Models



$$-log(p_\theta(x_0)) \leq \log(\frac{q(x_{1:T}|x_0)}{p_\theta(x_0,x_{1:T})}) = -\log(p(x_T)) + \sum_{t=1}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)})$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log(\frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}) + \log(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)})$$

We now apply the Bayes and condition it on x0 to avoid high variances

$$q(x_t|x_{t-1}) = \frac{q(x_{t-1}|x_t)q(x_t)}{q(x_{t-1})} => \frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log(\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}) + \log(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)})$$

Diffusion Models

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

Diffusion Models

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

Diffusion Models

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

Diffusion Models

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p_\theta(x_0|x_1)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{p(x_T)}\right)$$

Diffusion Models

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p_\theta(x_0|x_1)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{p(x_T)}\right)$$

We can then write it in DKL notation

$$= D_{KL}(q(x_T|x_0)||p(x_T)) + \sum_{t=2} T D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

Diffusion Models

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p_\theta(x_0|x_1)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{p(x_T)}\right)$$

We can then write it in DKL notation

$$= \cancel{D_{KL}(q(x_T|x_0)||p(x_T))} + \sum_{t=2} T D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

Diffusion Models

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p_\theta(x_0|x_1)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{p(x_T)}\right)$$

We can then write it in DKL notation

$$= \cancel{D_{KL}(q(x_T|x_0)||p(x_T))} + \sum_{t=2} T D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

And the 1st term can be ignored since it will have a very low value

Diffusion Models

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \sum_{t=2}^{T} \log\left(\frac{q(x_t|x_0)}{q(x_{t-1}|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p(x_T)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{q(x_1|x_0)}\right) + \log\left(\frac{q(x_1|x_0)}{p_\theta(x_0|x_1)}\right)$$

$$= -\log(p_\theta(x_0|x_1)) + \sum_{t=2}^{T} \log\left(\frac{q(x_{t-1}|x_t,x_0)}{p_\theta(x_{t-1}|x_t)}\right) + \log\left(\frac{q(x_T|x_0)}{p(x_T)}\right)$$

We can then write it in DKL notation

$$= \cancel{D_{KL}(q(x_T|x_0)||p(x_T))} + \sum_{t=2} T D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

And the 1st term can be ignored since it will have a very low value

$$\sum_{t=2} T D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

$$\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

$$\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t x_0}, (1-\bar{\alpha}_t)I)$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

$$\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t x_0}, (1-\bar{\alpha}_t)I)$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_t|x_{t-1}) = N(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

$$\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t x_0}, (1-\bar{\alpha}_t)I)$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \qquad x_t = \sqrt{\bar{\alpha}_t x_0} + \sqrt{1-\bar{\alpha}_t}$$

Diffusion Models

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$$

$$x_t = \sqrt{\bar{\alpha}_t x_0} + \sqrt{1-\bar{\alpha}_t}$$

We can now use bayes theorem to find the posterior $q(x_{t-1}|x_t, x_0)$ in terms Of $\tilde{\beta}_t$ and $\tilde{\mu}_t(x_t, x_0)$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t(1-\bar{\alpha}_{t-1})}}{1-\bar{\alpha}_t}x_t$$

Diffusion Models

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \qquad x_t = \sqrt{\bar{\alpha}_t x_0} + \sqrt{1-\bar{\alpha}_t}$$

We can now use bayes theorem to find the posterior $q(x_{t-1}|x_t, x_0)$ in terms Of $\tilde{\beta}_t$ and $\tilde{\mu}_t(x_t, x_0)$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t(1-\bar{\alpha}_{t-1})}}{1-\bar{\alpha}_t}x_t$$

Diffusion Models

$$x_t = \sqrt{\bar{\alpha}_t x_0} + \sqrt{1 - \bar{\alpha}_t}$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t(1-\bar{\alpha}_{t-1})}}{1-\bar{\alpha}_t} x_t$$

We can now try to predict x0 changing xt by x0 then ending up with the following

$$\mu\ (x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon\ (x_t, t) \right)$$

Diffusion Models

Going back to what we had for our result we must try to find our loss function that we will be simply using a MSE between the actual μ and the predicted μ

$$L_t = \frac{1}{2\sigma^2} ||\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)||^2$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)$$

$$\tilde{\mu}_\theta(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)$$

$$\tilde{\mu}_\theta(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)$$

$$\tilde{\mu}_\theta(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$$

$$L_t = \frac{1}{2\sigma^2}||\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)||^2$$

Diffusion Models

$$\sum_{t=2} TD_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) - \log(p_\theta(x_0|x_1))$$

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I)$$

$$p(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)$$

$$\tilde{\mu}_\theta(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right)$$

$$L_t = \frac{\beta_t^2}{2\sigma^2\alpha_t(1-\bar{\alpha}_t)}||\epsilon - \epsilon_\theta(x_t, t)||^2$$

Diffusion Models

The author decided to ignore the scheduling term at the front to simplify the implementation as well as obtain better sampling

$$L_t = \frac{\beta_t^2}{2\sigma^2\alpha_t(1-\bar{\alpha}_t)}||\epsilon - \epsilon_\theta(x_t, t)||^2$$

Diffusion Models

The author decided to ignore the scheduling term at the front to simplify the implementation as well as obtain better sampling

$$L_t = \frac{\beta_t^2}{2\sigma^2 \alpha_t (1-\bar{\alpha}_t)} ||\epsilon - \epsilon_\theta(x_t, t)||^2$$

Diffusion Models

The author decided to ignore the scheduling term at the front to simplify the implementation as well as obtain better sampling

$$L_t = \frac{\beta_t^2}{2\sigma^2\alpha_t(1-\bar{\alpha}_t)}||\epsilon - \epsilon_\theta(x_t, t)||^2$$

We now end up with a term that is basically the actual noise at t and the predicted noise at t by our neural network given an image

$$L_t = ||\epsilon - \epsilon_\theta(x_t, t)||^2$$

Diffusion Models

The author decided to ignore the scheduling term at the front to simplify the implementation as well as obtain better sampling

$$L_t = \frac{\beta_t^2}{2\sigma^2\alpha_t(1-\bar{\alpha}_t)}||\epsilon - \epsilon_\theta(x_t, t)||^2$$

We now end up with a term that is basically the actual noise at t and the predicted noise at t by our neural network given an image

$$L_t = ||\epsilon - \epsilon_\theta(x_t, t)||^2$$

Diffusion Models

We now have one last term to take care of and that is the last term in the Lower bound $L = \sum_{t=2}^{T} ||\epsilon - \epsilon_\theta(x_t, t)||^2 - \log(p_\theta(x_0|x_1))$

Since this value will result in a product of an integral for each pixel of the image the authors decided to also ignore this term thus ending with out final Loss function of

$$L_{simple} = E_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

Diffusion Models

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on

$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$

6: **until** converged

Diffusion Models

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \mathbf{z}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

$$p_\theta(\mathbf{x}_0|\mathbf{x}_1) = \prod_{i=1}^{D} \int_{\delta_-(x_0^i)}^{\delta_+(x_0^i)} \mathcal{N}(x; \mu_\theta^i(\mathbf{x}_1, 1), \beta_1)\, dx$$

$$\delta_+(x) = \begin{cases} \infty & \text{if } x = 1 \\ x + \frac{1}{255} & \text{if } x < 1 \end{cases} \qquad \delta_-(x) = \begin{cases} -\infty & \text{if } x = -1 \\ x - \frac{1}{255} & \text{if } x > -1 \end{cases}$$

Coding a Diffusion Model :)

Thank you :)