Generative Models

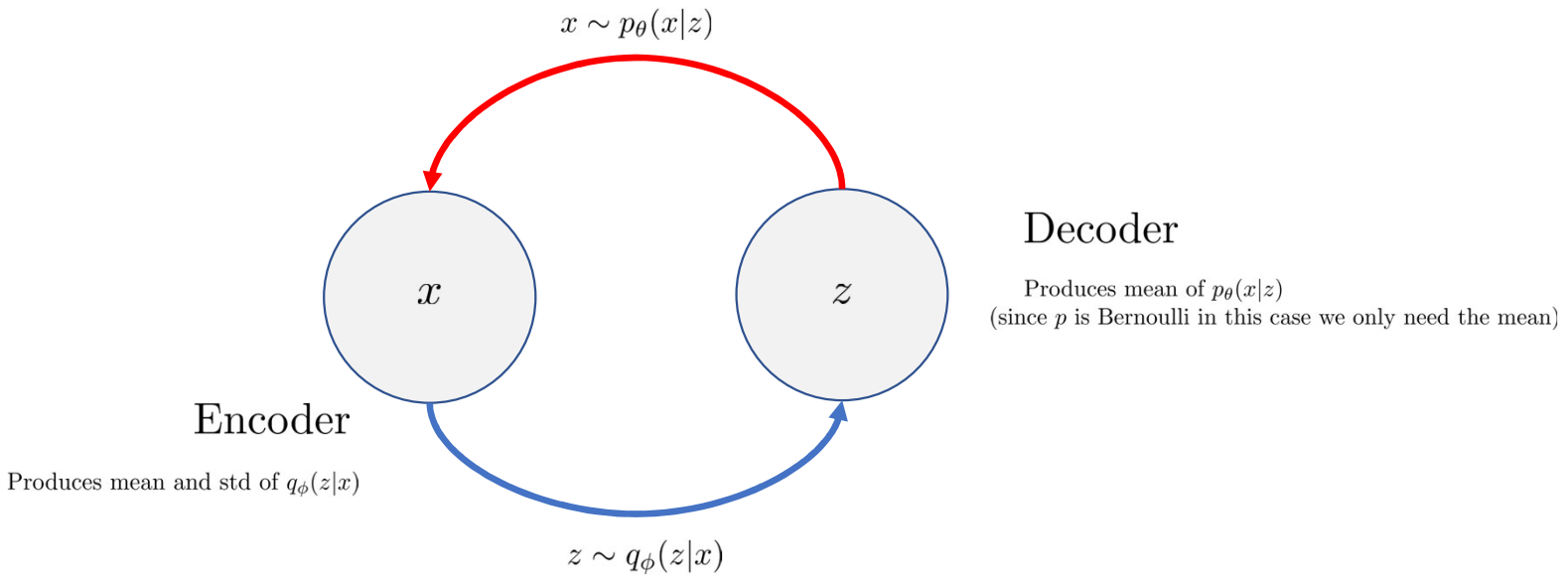
Lecture 8

Diffusion Models

Diffusion models are at the moment the State Of The Art in generative models. They are the key ingredients in prompt based image generation like DALL-E and Stable Diffusion.

Recall the VAE model.

We can visualize it by a diagram



Now we can extend the VAE model. Instead of just one latent variable we can consider a sequence

$$\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T$$

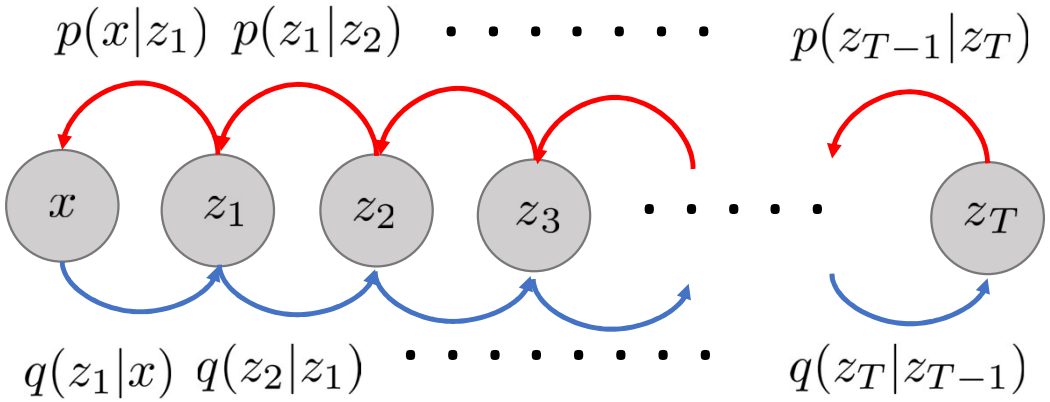
with conditional distributions

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots, \mathbf{z}_1, \mathbf{x})$$

We shall only consider Markovian models where

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{z}_{t-2}, \ldots, \mathbf{z}_1, \mathbf{x}) = q(\mathbf{z}_t|\mathbf{z}_{t-1})$$

A *Hierachical VAE* could be described by a diagram



where each of the arrows are Neural Networks (red=Decoder, blue=Encoder)

Using the Markov property we can write

$$
\begin{aligned}
q_\phi(\mathbf{z}_{1:T}|\mathbf{x}) &= q_\phi(\mathbf{z}_T, \mathbf{z}_{T-1}, \mathbf{z}_{T-2}, \ldots, \mathbf{z}_1, \mathbf{x}) / q_\phi(\mathbf{x}) \\
&= q_\phi(\mathbf{z}_{T-1}, \mathbf{z}_{T-2}, \ldots, \mathbf{z}_1, \mathbf{x}) q_\phi(\mathbf{z}_T|\mathbf{z}_{T-1}, \mathbf{z}_{T-2}, \ldots, \mathbf{z}_1, \mathbf{x}) / q_\phi(\mathbf{x}) \\
&= q_\phi(\mathbf{z}_{T-1}, \mathbf{z}_{T-2}, \ldots, \mathbf{z}_1, \mathbf{x}) q_\phi(\mathbf{z}_T|\mathbf{z}_{T-1}) / q_\phi(\mathbf{x}) \\
&= q_\phi(\mathbf{z}_T|\mathbf{z}_{T-1}) q_\phi(\mathbf{z}_{T-1}, \mathbf{z}_{T-2}, \ldots, \mathbf{z}_1, \mathbf{x}) / q_\phi(\mathbf{x}) \\
&= q_\phi(\mathbf{z}_T|\mathbf{z}_{T-1}) q_\phi(\mathbf{z}_{T-1}, \mathbf{z}_{T-2}, \ldots, \mathbf{z}_1|\mathbf{x}) \\
&\vdots \\
&= q_\phi(\mathbf{z}_T|\mathbf{z}_{T-1}) q_\phi(\mathbf{z}_{T-1}|\mathbf{z}_{T-2}) \ldots q_\phi(\mathbf{z}_1, x) / q_\phi(\mathbf{x}) \\
&= q_\phi(\mathbf{z}_1|\mathbf{x}) \prod_{t=2}^{T} q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1})
\end{aligned}
$$

Requiring the Decoder to be Markov as well, we get

$$
\begin{aligned}
p_\theta(\mathbf{x}, \mathbf{z}_{1:T}) &= p_\theta(\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T) p_\theta(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T) \\
&= p_\theta(\mathbf{x}|\mathbf{z}_1) p_\theta(\mathbf{z}_1|\mathbf{z}_2, \ldots, \mathbf{z}_T) p_\theta(\mathbf{z}_2, \ldots, \mathbf{z}_T) \\
&= p_\theta(\mathbf{x}|\mathbf{z}_1) p_\theta(\mathbf{z}_1|\mathbf{z}_2) p_\theta(\mathbf{z}_2, \ldots, \mathbf{z}_T) \\
&\vdots \\
&= p_\theta(\mathbf{x}|\mathbf{z}_1) p_\theta(\mathbf{z}_1|\mathbf{z}_2) \ldots p_\theta(\mathbf{z}_{T-1}|\mathbf{z}_T) p_\theta(\mathbf{z}_T) \\
&= p_\theta(\mathbf{x}|\mathbf{z}_1) p_\theta(\mathbf{z}_T) \prod_{t=2}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)
\end{aligned}
$$

How would we train such a model?

As usual we would do Maximum Likelihood, parametrizing the encoder and decoder by Neural Networks ϕ and θ

$$\max_{\theta} \log p_\theta(\mathbf{x})$$

We can then write

$$
\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}, \mathbf{z}_{1:T}) d\mathbf{z}_{1:T} \\
&= \log \int \frac{p_\theta(\mathbf{x}, \mathbf{z}_{1:T})}{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})} q_\phi(\mathbf{z}_{1:T}|\mathbf{x}) d\mathbf{z}_{1:T} \\
&= \log \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})} \left(\frac{p_\theta(\mathbf{x}, \mathbf{z}_{1:T})}{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})} \right) \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})} \left(\log \frac{p_\theta(\mathbf{x}, \mathbf{z}_{1:T})}{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})} \right) \\
&= \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})} \left(\log \frac{p_\theta(\mathbf{x}|\mathbf{z}_1) p_\theta(\mathbf{z}_T) \prod_{t=2}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t))}{q_\phi(\mathbf{z}_1|\mathbf{x}) \prod_{t=2}^{T} q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1})} \right)
\end{aligned}
$$

The last expression we can split up as

$$\mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\log \frac{p_\theta(\mathbf{z}_T)p_\theta(\mathbf{x}|\mathbf{z}_1))}{q_\phi(\mathbf{z}_1|\mathbf{x})}\right) + \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|x)}\left(\sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1})}\right)$$

Now

$$q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{x}) = \frac{q_\phi(\mathbf{z}_t,\mathbf{z}_{t-1},\mathbf{x})}{q_\phi(\mathbf{z}_{t-1},\mathbf{x})} = \frac{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})q_\phi(\mathbf{z}_t,\mathbf{x})}{q_\phi(\mathbf{z}_{t-1},\mathbf{x})}$$

But

$$q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1},\mathbf{x}) = q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1})$$

if $t \geq 2$ by the Markovian property so we get

$$q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1}) = \frac{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})q_\phi(\mathbf{z}_t,\mathbf{x})}{q_\phi(\mathbf{z}_{t-1},\mathbf{x})}$$

It follows that

$$\sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1})} = \sum_{t=2}^{T} \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})} - \log q_\phi(\mathbf{z}_t,\mathbf{x}) + \log q_\phi(\mathbf{z}_{t-1},\mathbf{x}) \right)$$

The sum

$$\sum_{t=2}^{T} - \log q_\phi(\mathbf{z}_t, \mathbf{x}) + \log q_\phi(\mathbf{z}_{t-1}, \mathbf{x})$$

telescopes to just

$$\begin{aligned}
\log q_\phi(\mathbf{z}_1, \mathbf{x}) - \log q_\phi(\mathbf{z}_T, \mathbf{x}) &= \log q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(\mathbf{x}) - \log q_\phi(\mathbf{z}_T|\mathbf{x})q_\phi(\mathbf{x}) \\
&= \log q_\phi(\mathbf{z}_1|\mathbf{x})) + \log q_\phi(\mathbf{x}) - (\log q_\phi(\mathbf{z}_T|\mathbf{x}) + \log q_\phi(\mathbf{x})) \\
&= \log q_\phi(\mathbf{z}_1|\mathbf{x}) - \log q_\phi(\mathbf{z}_T|\mathbf{x})
\end{aligned}$$

Taking expectations we get

$$\mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\log \frac{p_\theta(\mathbf{z}_T)p_\theta(\mathbf{x}|\mathbf{z}_1)}{q_\phi(\mathbf{z}_1|\mathbf{x})}\right) + \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\sum_{t=2}^{T}\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_t|\mathbf{z}_{t-1})}\right)$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\log \frac{p_\theta(\mathbf{z}_T)p_\theta(\mathbf{x}|\mathbf{z}_1))}{q_\phi(\mathbf{z}_1|\mathbf{x})}\right) + \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\sum_{t=2}^{T}\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right)$$

$$+ E_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}(\log q_\phi(\mathbf{z}_1|\mathbf{x}) - \log q_\phi(\mathbf{z}_T|\mathbf{x}))$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}(\log p(\mathbf{z}_T)) + \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}(p_\theta(\mathbf{x}|\mathbf{z}_1)) - \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}(\log q_\phi(\mathbf{z}_1|\mathbf{x}))$$

$$+ \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\sum_{t=2}^{T}\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right) + E_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}(\log q_\phi(\mathbf{z}_1|\mathbf{x}) - \log q_\phi(\mathbf{z}_T|\mathbf{x}))$$

Collecting terms and rearranging we get

$$\mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}(p_\theta(\mathbf{x}|\mathbf{z}_1)) + \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\sum_{t=2}^{T}\log\frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right)$$

$$+ \mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\log\frac{p_\theta(\mathbf{z}_T)}{q_\phi(\mathbf{z}_T|\mathbf{x})}\right)$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}(p_\theta(\mathbf{x}|\mathbf{z}_1)) + \sum_{t=2}^{T}\mathbb{E}_{q_\phi(\mathbf{z}_{t-1},\mathbf{z}_t|\mathbf{x})}\left(\log\frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right)$$

$$+ \mathbb{E}_{q_\phi(\mathbf{z}_T|\mathbf{x})}\left(\frac{p_\theta(\mathbf{z}_T)}{q_\phi(\mathbf{z}_T|\mathbf{x})}\right)$$

Here we have used that

$$
\begin{aligned}
&\mathbb{E}_{q_\phi(\mathbf{z}_{1:T}|\mathbf{x})}\left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right) \\
&= \int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right) q_\phi(\mathbf{z}_{1:T}|\mathbf{x}) d\mathbf{z}_{1:T} \\
&= \int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right) q_\phi(\mathbf{z}_T,\mathbf{z}_{T-1},\ldots,\mathbf{z}_1|\mathbf{x}) d\mathbf{z}_T d\mathbf{z}_{T-1}\ldots d\mathbf{z}_1 \\
&= \int \left(\int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right) q_\phi(\mathbf{z}_T,\mathbf{z}_{T-1},\ldots,\mathbf{z}_1|\mathbf{x}) d\mathbf{z}_T\ldots d\mathbf{z}_{t+1}\right) d\mathbf{z}_t\ldots d\mathbf{z}_1 \\
&= \int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})}\right) \left(\int q_\phi(\mathbf{z}_T,\mathbf{z}_{T-1},\ldots,\mathbf{z}_1|\mathbf{x}) d\mathbf{z}_T\ldots d\mathbf{z}_{t+1}\right) d\mathbf{z}_t\ldots d\mathbf{z}_1
\end{aligned}
$$

By marginalization

$$\int q_\phi(\mathbf{z}_T, \mathbf{z}_{T-1}, \ldots, \mathbf{z}_1|\mathbf{x}) d\mathbf{z}_T \ldots d\mathbf{z}_{t+1} = q_\phi(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_1|\mathbf{x})$$

so

$$\int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})} \right) \left(\int q_\phi(\mathbf{z}_T, \mathbf{z}_{T-1}, \ldots, \mathbf{z}_1|\mathbf{x}) d\mathbf{z}_T \ldots d\mathbf{z}_{t+1} \right) d\mathbf{z}_t \ldots d\mathbf{z}_1$$
$$= \int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})} \right) q_\phi(\mathbf{z}_t, \mathbf{z}_{t-1}, \ldots, \mathbf{z}_1|\mathbf{x}) d\mathbf{z}_t d\mathbf{z}_{t-1} \ldots d\mathbf{z}_1$$

By the same marginalization argument this is

$$\int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})} \right) q_\phi(\mathbf{z}_t, \mathbf{z}_{t-1}|\mathbf{x}) d\mathbf{z}_t d\mathbf{z}_{t-1}$$

Now

$$q_\phi(\mathbf{z}_t, \mathbf{z}_{t-1}|\mathbf{x}) = \frac{q_\phi(\mathbf{z}_t, \mathbf{z}_{t-1}, \mathbf{x})}{q_\phi(\mathbf{x})} = \frac{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})q_\phi(\mathbf{z}_t, \mathbf{x})}{q_\phi(\mathbf{x})} = q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})q_\phi(\mathbf{z}_t|\mathbf{x})$$

So

$$\begin{aligned}
&\int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})}\right) q_\phi(\mathbf{z}_t, \mathbf{z}_{t-1}|\mathbf{x}) d\mathbf{z}_t d\mathbf{z}_{t-1} \\
&= \int \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})}\right) q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) q_\phi(\mathbf{z}_t|\mathbf{x}) d\mathbf{z}_{t-1} d\mathbf{z}_t \\
&= \int \mathbb{E}_{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})} \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})}\right) q_\phi(\mathbf{z}_t|\mathbf{x}) d\mathbf{z}_t \\
&= -\mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x})}(D_{KL}(q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})||p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t))
\end{aligned}$$

So the $ELBO$ for the Hierachical VAE is

$$
\begin{aligned}
ELBO \\
&= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}(p_\theta(\mathbf{x}|\mathbf{z}_1)) + \sum_{t=2}^{T} \mathbb{E}_{q_\phi(\mathbf{z}_{t-1},\mathbf{z}_t|\mathbf{x})} \left(\log \frac{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}{q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})} \right) \\
&+ \mathbb{E}_{q_\phi(\mathbf{z}_T|\mathbf{x})} \left(\frac{p_\theta(\mathbf{z}_T)}{q_\phi(\mathbf{z}_T|\mathbf{x})} \right) \\
&= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}(p_\theta(\mathbf{x}|\mathbf{z}_1)) - \sum_{t=2}^{T} \mathbb{E}_{q_\phi(\mathbf{z}_t|\mathbf{x})} \left(D_{KL}(q_\phi(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})||p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)) \right. \\
&- D_{KL}(q_\phi(\mathbf{z}_T|\mathbf{x})||p_\theta(\mathbf{z}_T))
\end{aligned}
$$

This formula is very much like the $ELBO$ for a regular VAE

$$ELBO = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}(p_\theta(\mathbf{x}|\mathbf{z})) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_{prior}(\mathbf{z}))$$

The *reconstruction term* appears in both formulas, the middle term measures the KL-divergence between the decoder distribution and the *reverse* encoder distribution

The *Denoising Diffusion Probabilistic Model (DDPM)* uses a deterministic Encoder.

The model inputs an image (or other type of data point). We then sequentially add Gaussian noise to the image over many steps (~ 1000) to turn the image into *white noise*

An image is a tensor of dimension $C \times H \times W$ where C is the number of channels for each pixel. So for RGB images $C = 3$

Assume we are given a data set, let's say of images just the fix ideas but it could be other types as well.

$$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$$

we assume these are samples from some distribution data distribution q_0 of dimension $m = C \times H \times W$

As usual our goal is to generate samples from the data distribution

We are going to slowly add Gaussian noise to the image

Assume $\mathbf{x}$ is the original image and we have done t steps to a noisy image $\mathbf{z}_t$.

- The dimension of the latent variables are equal to the dimension of the data i.e. $\dim \mathbf{x} = \dim \mathbf{z}_t$ for all t
- The conditional distributions $q(\mathbf{z}_{t+1}|\mathbf{z}_t)$ distributions are fixed i.e. the do not depend on a set of learnable parameters. It is a Gaussian centered around $\mathbf{z}_t$ and a diagonal covariance matrix, which can vary with t
- The distribution of the final latent variable, when we have added a sufficient amount of noise, $\mathbf{z}_T$ is a sample from a standard Gaussian $\mathcal{N}(\underline{0}, I)$

We parametrize the distributions by

$$q(\mathbf{z}_{t+1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t+1}|\sqrt{\alpha_t}\mathbf{z}_t, (1-\alpha_t)I)$$

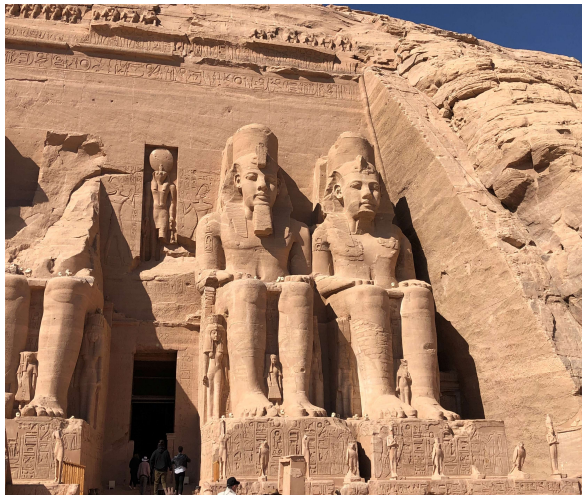
Thus the sequence of random variables $\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T$ is a Markov chain with transition distributions $q(\mathbf{z}_{t+1}|\mathbf{z}_t) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{z}_t, (1-\alpha_t)I)$

It follows that

$$\mathbf{z}_{t+1} = \sqrt{\alpha_t}\mathbf{z}_t + \sqrt{(1-\alpha_t)}\varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, I)$

Let's start with an image $\mathbf{x}$. This is a tensor of dimension $(3, 600, 600)$

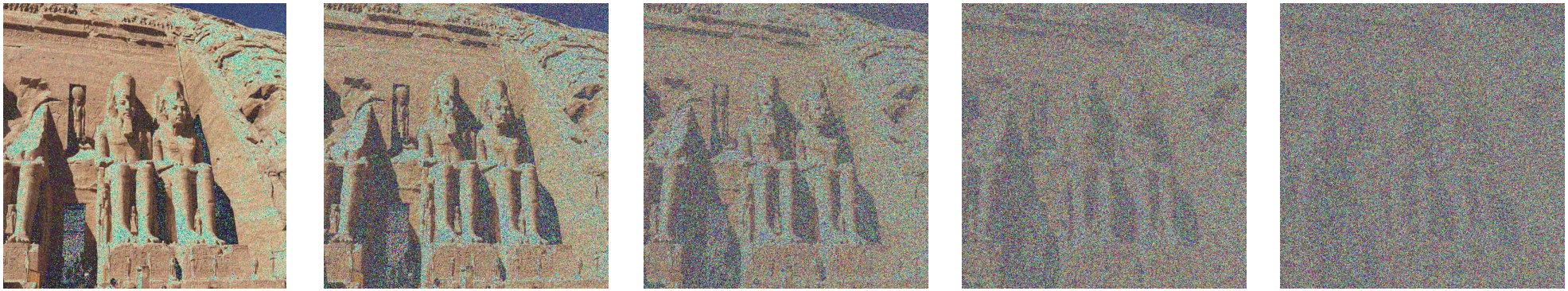


We put $\alpha_1 = 0.99$. Then we can can then draw a sample from $q(\mathbf{z}_1|\mathbf{x})$ by

$$\mathbf{z}_1 = \sqrt{\alpha_1}\mathbf{x} + \sqrt{1-\alpha_1}\varepsilon$$

where $\underline{\varepsilon}$ is a $(3, 600, 600)$ tensor of samples from $\mathcal{N}(0, 1)$

z_1 z_2 z_3 z_4 z_5



$\alpha_1 = 0.99$ $\alpha_2 = 0.98$ $\alpha_3 = 0.97$ $\alpha_4 = 0.96$ $\alpha_5 = 0.95$

If we could reverse the diffusion steps i.e. if we could compute the 'backwards' process $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$, we could start with $\mathbf{z}_T$ which is 'white noise' i.e. $\mathbf{z}_T \sim \mathcal{N}(0, I)$ and then step by step sample from $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$ we would eventually sample from $q(\mathbf{x}|\mathbf{z}_1)$.

It is known that if the variance $(1-\alpha_t)$ is sufficiently small then the 'reverse transition', $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$ is also Gaussian. Unfortunately the reverse process is intractable.

$$q(\mathbf{z}_1|\mathbf{x}) = \mathcal{N}(\mathbf{z}_1|\sqrt{\alpha_1}\mathbf{x}, (1-\alpha_1)\, I)$$

so

$$\mathbf{z}_1 = \sqrt{\alpha_1}\mathbf{x} + \sqrt{1-\alpha_1}\,\varepsilon$$

with

$$\varepsilon \sim \mathcal{N}(0, I) \ldots \text{ an } m - \textit{dimensional} \text{ vector}$$

The density function of a sum of random variables is the convolution of the density functions, so if $q_0(\mathbf{x})$ is the true data distribution, the density function of $\mathbf{z}_1$ is the convolution of $q_0(\sqrt{\alpha_1}\mathbf{x})$ with $\mathcal{N}(0, (1-\alpha_0)I)$ i.e.

$$q_1(\mathbf{z}_1) = \int \frac{1}{\sqrt{\alpha_1}^m} q_0(\sqrt{\alpha}_1\mathbf{x})\mathcal{N}(\mathbf{z}_1 - \alpha_1\mathbf{x}|0, (1-\alpha_1)I)d\mathbf{x}$$

and at each step

$$q_t(\mathbf{z}_t) = \int \frac{1}{\sqrt{\alpha_t}^m} q_{t-1}(\sqrt{\alpha_t}\mathbf{z}_{t-1})\mathcal{N}(\mathbf{z}_t - \sqrt{\alpha_t}\mathbf{z}_{t-1}|0, (1-\alpha_t)I)d\mathbf{z}_{t-1}$$

We have

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_{t-1} + \sqrt{1-\alpha_t}\,\varepsilon_t$$

where $\varepsilon_t \sim \mathcal{N}(\underline{0}, I)$ (an m-dimensional vector) and

$$\mathbf{z}_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{z}_{t-2} + \sqrt{1-\alpha_{t-1}}\,\underline{\varepsilon}_{t-1}$$

Thus

$$\begin{aligned}
\mathbf{z}_t &= \sqrt{\alpha_t}\mathbf{z}_{t-1} + \sqrt{1-\alpha_t}\,\varepsilon_t \\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\mathbf{z}_{t-2} + \sqrt{1-\alpha_{t-1}}\,\varepsilon_{t-1}) + \sqrt{1-\alpha_t}\,\varepsilon_t \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{z}_{t-2} + (\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\,\varepsilon_{t-1} + \sqrt{1-\alpha_t}\,\varepsilon_t)
\end{aligned}$$

The sum of the two normal random vaiables is normal with mean equal to the sum and variance equal to the sum of the variances. It follows that

$$\begin{aligned}
&(\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\,\varepsilon_{t-1} + \sqrt{1-\alpha_t}\,\varepsilon_t) \\
\sim \mathcal{N}(\underline{0}, (\alpha_t - \alpha_t\alpha_{t-1} + (1-\alpha_t))I) &= N(0, (1-\alpha_t\alpha_{t-1})\,I)
\end{aligned}$$

so

$$\mathbf{z}_t = \sqrt{\alpha_t\alpha_{t-1}}\mathbf{z}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\,\varepsilon$$

It follows that if we put

$$\overline{\alpha}_t = \alpha_t \alpha_{t-1} \ldots \alpha_0$$

then

$$\mathbf{z}_t = \sqrt{\overline{\alpha}_t}\, \mathbf{x} + \sqrt{1 - \overline{\alpha}_t}\, \overline{\varepsilon}_t$$

and

$$q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\mathbf{z}_t|\sqrt{\overline{\alpha}_t}\, \mathbf{x}, (1 - \overline{\alpha}_t)\, I)$$

The backwards process $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$ is intractable but it turns out that if we also condition on $\mathbf{x}$, the distributions $q(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x})$ are tractable

We have

$$
\begin{aligned}
q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) &= \frac{q(\mathbf{z}_t, \mathbf{z}_{t-1}, \mathbf{x})}{q(\mathbf{z}_t, \mathbf{x})} \\
&= \frac{q(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{x})q(\mathbf{z}_{t-1}, \mathbf{x})}{q(\mathbf{z}_t, \mathbf{x})} \\
&= \frac{q(\mathbf{z}_t|\mathbf{z}_{t-1})q(\mathbf{z}_{t-1}, \mathbf{x})}{q(\mathbf{z}_t, \mathbf{x})} \ldots \text{ Markov} \\
&= \frac{q(\mathbf{z}_t|\mathbf{z}_{t-1})q(\mathbf{z}_{t-1}|\mathbf{x})}{q(\mathbf{z}_t|\mathbf{x})} \ldots \text{ dividing numerator and denominator with } q(\mathbf{x})
\end{aligned}
$$

It follows that

$$q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) = \frac{\mathcal{N}(\mathbf{z}_t|\sqrt{\alpha_t}\mathbf{z}_{t-1}, (1-\alpha_t)I)\mathcal{N}(\mathbf{z}_{t-1}|\sqrt{\overline{\alpha}_{t-1}}\,\mathbf{x}, (1-\overline{\alpha}_{t-1})\,I)}{\mathcal{N}(\mathbf{z}_t|\sqrt{\overline{\alpha}_t}\,\mathbf{x}, (1-\overline{\alpha}_t)\,I)}$$

A long computation shows that

$$q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) = \mathcal{N}(\mathbf{z}_{t-1}|\mu(\mathbf{z}_t, \mathbf{x}), \Sigma(\mathbf{z}_t, \mathbf{x}))$$

with

$$\mu(\mathbf{z}_t, \mathbf{x}) = \left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)}\right)\mathbf{z}_t + \left(\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{(1-\overline{\alpha}_t)}\right)\mathbf{x}$$

and

$$\Sigma(\mathbf{z}_t, \mathbf{x}) = \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)} \cdot I$$

Since the reverse process is intractable we want to approximate by decoders $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$ where θ is a Neural Net.

As we showed the $ELBO$ is

$$\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})}(p_\theta(\mathbf{x}|\mathbf{z}_1)) - \sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}\left(D_{KL}(q(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})||p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)\right)$$
$$- D_{KL}(q(\mathbf{z}_T|\mathbf{x})||p_\theta(\mathbf{z}_T))$$

We have

$$q(\mathbf{z}_T|\mathbf{x}) = \mathcal{N}(\mathbf{z}_T|\sqrt{\overline{\alpha_T}}\mathbf{x}, (1-\overline{\alpha_T})I)$$

We let $\alpha_t \to 0$ so $\overline{\alpha_T} \simeq 0$ and

$$q(\mathbf{z}_T|\mathbf{x}) \simeq \mathcal{N}(0, I)$$

We also let $p_\theta(\mathbf{z}_T) = \mathcal{N}(0, I)$ so the term $D_{KL}(q(\mathbf{z}_T|\mathbf{x})||p_\theta(\mathbf{z}_T)) = 0$

Since $q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})$ are Gaussians we also want the distributions $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$ to be Gaussians with some $mean = \mu(\theta, \mathbf{z}_t)$ and $covariance = \sigma(\theta, \mathbf{z}_t)I$

We will use the following general fact

$$\underset{\mu,\Sigma}{\operatorname{argmin}}\ D_{KL}(q(z)||\mathcal{N}(z|\mu, \Sigma)) = mean(q), variance(q)$$

So the parameters $\mu(\mathbf{z}_t)$ and $\sigma(\mathbf{z}_t)$ that minimize $D_{KL}(q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})||p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t))$ are

$$\mu(\mathbf{z}_t) = mean(q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x})) = \left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)}\right)\mathbf{z}_t + \left(\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{(1-\overline{\alpha}_t)}\right)\mathbf{x}$$

and

$$\sigma_t = \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)} \ldots \text{ does not depend on } \mathbf{z}_t$$

We have

$$\mathbf{z}_t = \sqrt{\overline{\alpha}_t}\,\mathbf{x} + \sqrt{1-\overline{\alpha}_t}\,\underline{\varepsilon}$$

so

$$\mathbf{x} = \frac{1}{\sqrt{\overline{\alpha}_t}}\left(\mathbf{z}_t - \sqrt{1-\overline{\alpha}_t}\,\overline{\varepsilon}_t\right)$$

Now we put this into the expression for $\mu(\mathbf{z}_t)$

$$
\begin{aligned}
\mu(\mathbf{z}_t) &= \left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)}\right)\mathbf{z}_t + \left(\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{(1-\overline{\alpha}_t)}\right)\mathbf{x} \\
&= \left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)}\right)\mathbf{z}_t + \left(\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{(1-\overline{\alpha}_t)}\right)\frac{1}{\sqrt{\overline{\alpha_t}}}\left(\mathbf{z}_t - \sqrt{1-\overline{\alpha_t}}\,\overline{\varepsilon_t}\right) \\
&= \left(\left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)}\right) + \left(\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{(1-\overline{\alpha}_t)}\right)\frac{1}{\sqrt{\overline{\alpha_t}}}\right)\mathbf{z}_t \\
&- \left(\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{(1-\overline{\alpha}_t)}\right)\frac{1}{\sqrt{\overline{\alpha_t}}}\sqrt{1-\overline{\alpha_t}}\,\overline{\varepsilon_t} \\
&= \frac{\alpha_t(1-\overline{\alpha_{t-1}}) + (1-\alpha_t)}{\sqrt{\alpha_t}(1-\alpha_t)}\mathbf{z}_t - \frac{(1-\alpha_t)}{\sqrt{\alpha_t}\sqrt{(1-\overline{\alpha_t})}}\,\overline{\varepsilon_t} \\
&= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{(1-\alpha_t)}{\sqrt{(1-\overline{\alpha_t})}}\overline{\varepsilon_t}\right)
\end{aligned}
$$

So $\mu(\mathbf{z}_t)$ becomes our target i.e. we want to find θ such that

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}|\mu(\theta, \mathbf{z}_t), \sigma(\theta, \mathbf{z}_t)I)$$

minimizes

$$\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}(||\mu(\theta, \mathbf{z}_t) - \mu(\mathbf{z}_t)||^2) \text{ for all } t$$

Since the latent variable $\mathbf{z}_t$ is given at training time, i.e. does not depend on θ we can write

$$\mu(\theta, \mathbf{z}_t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}} f_\theta(\mathbf{z}_t))$$

where $f_\theta(\mathbf{z}_t)$ is the output of the Neural Net

Thus the loss function becomes

$$\begin{aligned}
&\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}(||\mu(\theta,\mathbf{z}) - \mu(\mathbf{z}_t||^2) \\
&= \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}(||\frac{1}{\sqrt{\alpha_t}}(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}} f_\theta(\mathbf{z}_t)) - \frac{1}{\sqrt{\alpha_t}}(\mathbf{z}_t - \frac{(1-\alpha_t)}{\sqrt{(1-\overline{\alpha_t})}}\overline{\varepsilon_t}||^2) \\
&= \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}(||\overline{\varepsilon_t} - f_\theta(\mathbf{z}_t)||^2)
\end{aligned}$$

Since $\overline{\varepsilon_t} \sim \mathcal{N}(0, I)$ is the random part of $q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\sqrt{\overline{\alpha_t}}\,\mathbf{x}, (1-\overline{\alpha_t})\,I)$ the expectation over $q(\mathbf{z}_t|\mathbf{x})$ is of the form

$$\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}(||\overline{\varepsilon_t} - f_\theta(\mathbf{z}_t)||^2) = const. + \sqrt{1-\overline{\alpha_t}}\,\mathbb{E}_{\mathcal{N}(0,I)}(||\varepsilon - f_\theta(\mathbf{z}_t)||^2)$$

So our Decoder function f_θ has to minimize

$$\mathbb{E}_{\varepsilon \sim \mathcal{N}(0,I)}(||\varepsilon - f_\theta(\mathbf{z}_t)||^2) \text{ for } t = 2, \ldots, T$$

How we structure and train the Neural Net to minimize the loss function will be the topic for the next (and final lecture). But suppose we have trained to find a θ that minimizes the loss.

Then

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}|\mu(\theta, \mathbf{z_t}), \sigma_t)$$

where

$$\mu(\theta, \mathbf{z}_t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} f_\theta(\mathbf{z}_t)) \text{ and } \sigma_t = \frac{(1-\alpha_t)(1-\overline{\alpha}_{t-1})}{(1-\overline{\alpha}_t)}$$

We fix a sequence of α's

$$\alpha_1 > \alpha_2 > \cdots > \alpha_T$$

Start with an image $\mathbf{x}$ and add noise till $\mathbf{z}_T \sim \mathcal{N}(0, I)$

We let $p_\theta(\mathbf{z}_T) = \mathcal{N}(0, I)$

Sample $\mathbf{z}_T \sim \mathcal{N}(0, I)$ and then sample $\mathbf{z}_{T-1} \sim p_\theta(\mathbf{z}_{T-1}|\mathbf{z}_T)$.

Continue all the way down to a sample from $p_\theta(\mathbf{x}|\mathbf{z}_1)$

Home Work 4

Make a descending sequence of α's of length 500.

Chose your favorite photo and scale or crop it to a 600×600 pixels image $\mathbf{x}$

Use your sequence of α's to sequentially add noise to the image to generate increasingly noisy images $\mathbf{z}_t$

At the end estimate how close $\mathbf{z}_T$ is to 'white noise' by estimating

$$\mathbb{E}_{\varepsilon \sim \mathcal{N}(0,I)}(||\varepsilon - \mathbf{z}_T||^2)$$

(Generate a number of samples from $(0, I)$ and compute average of the squared differences)