Introduction to Diffusion Models

Huidong Xie
huidong.xie@yale.edu
Yale

Weijie Gan
weijie.gan@wustl.edu
WUSTL

Generative Model

Given samples of observed data x, the goal of generative models is to generate new samples that do not exist in the original observed data. Data x could be anything, including images, audio, text, etc. When training a generative model, we aim to train the model that approximates the data distribution $p(x)$.

To generate non-existent samples, we build the generative models by mapping a noise distribution $p_z(z)$ to data distribution $p(x)$ as $G_\theta(z)$, where θ is the parameters of the model. If a well-trained model G can approximate this mapping, we can then generate new instances in the data space by sampling new z from the noise distribution p_z and then feed into the trained network. To generate a wide variety of samples, we usually choose p_z as Gaussian distribution. The noise distributions z can be understood as latent space representations of observed samples x.

One of the most popular generative models is the Generative Adversarial Networks (GAN). GAN contains 2 networks, generator G and discriminator D. G models the data distribution $p(x)$ and aims to generate new samples from Gaussian noise. The network D takes either $G(z)$ or samples from the observed data x as input, and tries to discriminate if an input sample is from observed data x or is generated by network G.

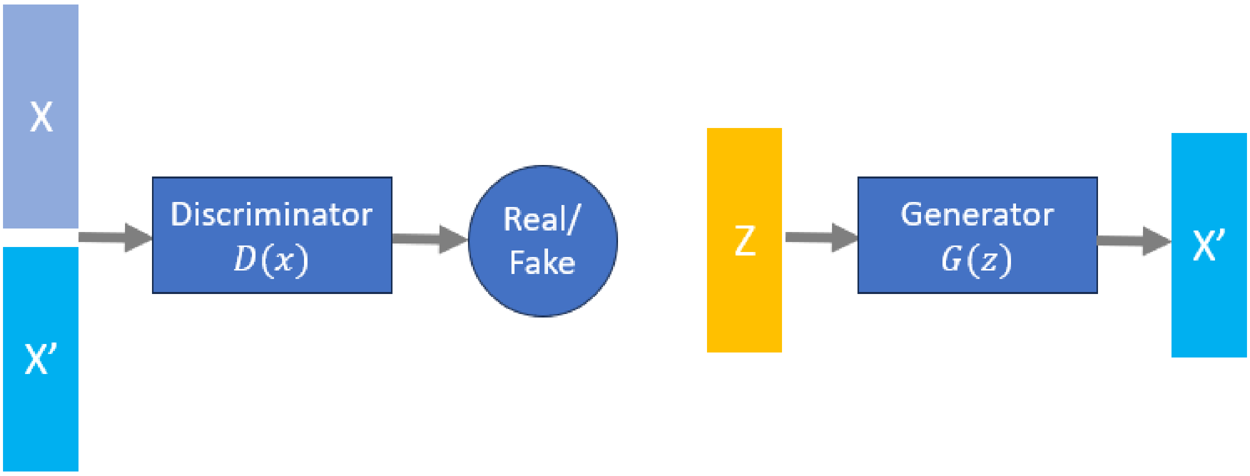

Figure 1: Simplified diagram of GAN.

GAN is an effective image synthesis method that produces promising results. Diffusion is another model that becomes increasingly popular recently for image synthesis. In this tutorial, we will start from the DDPM paper (denoising diffusion probabilistic model) [1]. This was the first paper demonstrating the use of diffusion model for synthesizing high-quality images.

Diffusion Model

In diffusion model, instead of sampling Gaussian noise as input to the generator network, we gradually add t steps of Gaussian noise to the images until the original information in the input is completely destroyed. This "gradually adding noise" process is called the forward process or diffusion process. This process is fixed to a Markov chain that gradually adds Gaussian noise.

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}),\ q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

where $\sqrt{1-\beta_t}$ and β_t represent the mean and variance of the Gaussian distribution at diffusion step t. In the DDPM paper, authors re-parameterize $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

A Markov chain or Markov process is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event. Specifically, it means that the image at time step (diffusion step) t only depends on the image at previous step $t-1$.

Similarly, the reverse process, or "removing noise process" could also be defined as a Markov chain with learned means and variances, starting from complete Gaussian noise at the last time step T, $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$.

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \;\; p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \textstyle\sum_\theta(x_t, t)) \tag{2}$$

where θ represents the trained parameters of the diffusion model. $\mu_\theta(x_t, t)$ and $\sum_\theta(x_t, t))$ are the learned mean and variance at each time step.

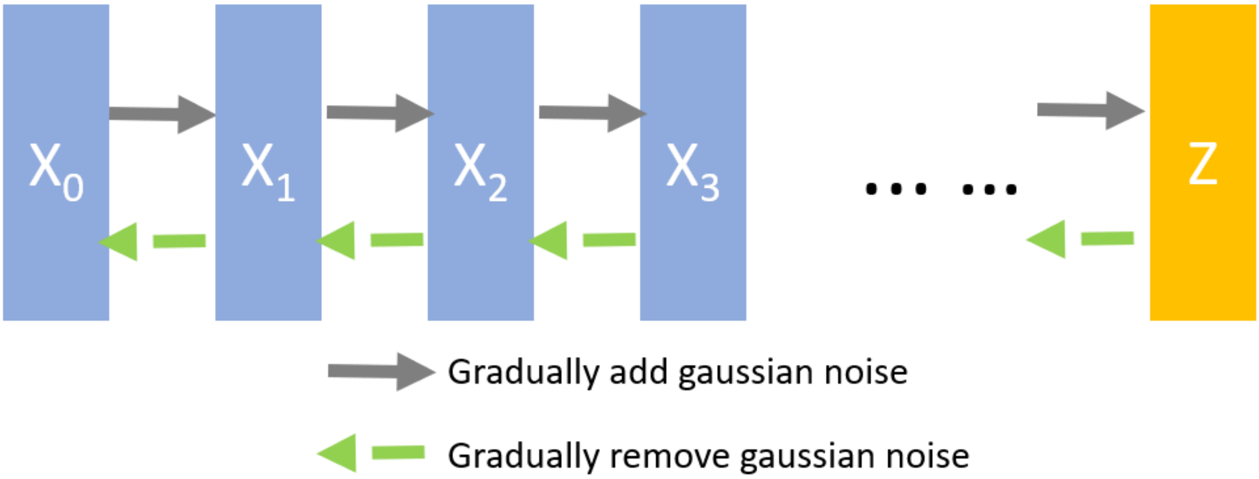

Figure 2: Simplified diagram of Diffusion models. x_0 is the same as x in Fig. 1.

We train the diffusion model so that we have the highest possibilities of observing original data distribution $(p_\theta(x_0))$. The network is trained by maximizing the log-likelihood of it (i.e., $\mathbb{E}(\log p_\theta(x_0)))$. In the DDPM paper, it says the training is performed by optimizing the evidence lower bound (ELBO or variational lower bound) of it. This is just a lower bound on the log-likelihood. Optimizing ELBO would be the same as optimizing $(-\log p_\theta(x_0))$.

$$\mathbb{E}(-\log p_\theta(x_0)) \leq \mathbb{E}_q\Big[-\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\Big] \tag{3}$$

The difference between ELOB and $\log p_\theta(x_0)$ is exactly the KL (Kullback–Leibler) divergence between $p_\theta(x_{0:T})$ and $q(x_{1:T}|x_0)$. KL divergence is a measure of the difference between 2 distributions.

$$\log p_\theta(x_0) = \log p_\theta(x_0) \int q(x_{1:T}|x_0) dx_{1:T} \qquad \text{(where } \int q(x_{1:T}|x_0) dx_{1:T} = 1) \tag{4}$$

$$= \int q(x_{1:T}|x_0)(\log p_\theta(x_0)) dx_{1:T} \tag{5}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}[\log p_\theta(x_0)] \qquad \text{(definition of Expectation)} \tag{6}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p_\theta(x_0, x_{1:T})}{p_\theta(x_{1:T}|x_0)}\Big] \qquad (p(x) = \frac{p(x,z)}{p(z|x)}) \tag{7}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p_\theta(x_0, x_{1:T}) q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0) q(x_{1:T}|x_0)}\Big] \tag{8}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p_\theta(x_0, x_{1:T})}{q(x_{1:T}|x_0)}\Big] + \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{1:T}|x_0)}\Big] \tag{9}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p_\theta(x_0, x_{1:T})}{q(x_{1:T}|x_0)}\Big] + KL(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0)) \qquad \text{(Definition of KL divergence)} \tag{10}$$

$$\geq \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p_\theta(x_0, x_{1:T})}{q(x_{1:T}|x_0)}\Big] \qquad \text{(KL is always larger than 0)} \tag{11}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\Big] \tag{12}$$

Through this derivation, we demonstrated the ELBO [2]. Then we can plug in Equations 1-2 into ELBO to get a more intuitive explanation of this objective.

$$\begin{aligned}
\log p_\theta(x_0) &\geq \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\Big] && (13)\\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}\Big] && (14)\\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)\prod_{t=2}^{T} p_\theta(x_{t-1}|x_t)}{q(x_T|x_{T-1})\prod_{t=1}^{T-1} q(x_t|x_{t-1})}\Big] && (15)\\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)\prod_{t=1}^{T-1} p_\theta(x_t|x_{t+1})}{q(x_T|x_{T-1})\prod_{t=1}^{T-1} q(x_t|x_{t-1})}\Big] && (16)\\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)}{q(x_T|x_{T-1})}\Big] + \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \prod_{t=1}^{T-1} \frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}\Big] && (17)\\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log p_\theta(x_0|x_1)\Big] + \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)}{q(x_T|x_{T-1})}\Big] + \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\sum_{t=1}^{T-1} \log \frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}\Big] && (18)\\
&= \mathbb{E}_{q(x_1|x_0)}\Big[\log p_\theta(x_0|x_1)\Big] + \mathbb{E}_{q(x_{T-1},x_T|x_0)}\Big[\log \frac{p(x_T)}{q(x_T|x_{T-1})}\Big] + \sum_{t=1}^{T-1} \mathbb{E}_{q(x_t,x_{t+1},x_{t-1}|x_0)}\Big[\log \frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}\Big] && (19)\\
&= \mathbb{E}_{q(x_1|x_0)}\Big[\log p_\theta(x_0|x_1)\Big] + \mathbb{E}_{q(x_{T-1}|x_0)}\Big[\mathbb{E}_{q(x_T)} \log \frac{p(x_T)}{q(x_T|x_{T-1})}\Big] + \\
&\quad \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t+1},x_{t-1}|x_0)}\Big[\mathbb{E}_{q(x_t)} \log \frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}\Big] && (20)\\
&= \mathbb{E}_{q(x_1|x_0)}\Big[\log p_\theta(x_0|x_1)\Big] - \mathbb{E}_{q(x_{T-1}|x_0)}\Big[KL(q(x_T|x_{T-1})||p(X_T))\Big] - \\
&\quad \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t+1},x_{t-1}|x_0)}\Big[KL(q(x_t|x_{t-1})||p_\theta(x_t|x_{t+1})))\Big] && (21)\\
&= \mathbb{E}_{q(x_1|x_0)}\Big[\log p_\theta(x_0|x_1)\Big] - \mathbb{E}_{q(x_{T-1}|x_0)}\Big[KL(q(x_T|x_{T-1})||p(x_T))\Big] \\
&\quad - \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t+1},x_{t-1}|x_0)}\Big[KL(q(x_t|x_{t-1})||p_\theta(x_t|x_{t+1})))\Big] && (22)
\end{aligned}$$

The first term $\mathbb{E}_{q(x_1|x_0)}\Big[\log p_\theta(x_0|x_1)\Big]$ is the log-likelihood of the original input x_0 given x_1, the image after the first diffusion step.

In the second term $\mathbb{E}_{q(x_{T-1}|x_0)}\Big[KL(q(x_T|x_{T-1})||p(x_T))\Big]$, $p(x_T)$ is completely Gaussian, and $q(x_{T-1}|x_0)$ is the last "adding noise" step. So this term is just difference between 2 Gaussian distribution, given the time step T is large enough. Since this is just the KL divergence between 2 Gaussian distributions, this term requires no training.

The third term $\sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t+1},x_{t-1}|x_0)}\Big[KL(q(x_t|x_{t-1})||p_\theta(x_t|x_{t+1})))\Big]$ is the difference between 2 distribution at each time step t.

To this end, we have a more intuitive explanation of the ELBO objective. You may notice that this objective (Equation 22) is different from the Equation (5) in the DDPM paper [1]. This is because the paper goes one step forward, rewriting the diffusion step as $q(x_t|x_{t-1}, x_0)$ instead of $q(x_t|x_{t-1})$. The disadvantage of using $q(x_t|x_{t-1})$ is that, this formula has 2 random variables x_t and x_{t-1}, which may be unstable in the training process. But luckily, we know x_0 (original input image) during training. Using Bayes rule, we can rewrite $q(x_t|x_{t-1}, x_0)$ and plug into ELBO.

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0) = \frac{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)} \tag{23}$$

$$\mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\Big] = \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}\Big] \tag{24}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)\prod_{t=2}^{T} p_\theta(x_{t-1}|x_t)}{q(x_1|x_0)\prod_{t=2}^{T} q(x_t|x_{t-1})}\Big] \tag{25}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)\prod_{t=2}^{T} p_\theta(x_{t-1}|x_t)}{q(x_1|x_0)\prod_{t=2}^{T} q(x_t|x_{t-1})}\Big] \tag{26}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)}{q(x_1|x_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}, x_0)}\Big] \tag{27}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)}{q(x_1|x_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(x_{t-1}|x_t)}{\frac{q(x_{t-1}|x_t,x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}}\Big] \tag{28}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)}{q(x_1|x_0)} + \log \prod_{t=2}^{T} \frac{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}\Big] \tag{29}$$

By expending the above equation for a few steps, you will notice that some terms just cancel out.

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)}{q(x_1|x_0)} + \log \frac{q(x_1|x_0)}{q(x_T|x_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\Big] \tag{30}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \Big(\frac{p(x_T)p_\theta(x_0|x_1)}{q(x_1|x_0)}\frac{q(x_1|x_0)}{q(x_T|x_0)}\Big) + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\Big] \tag{31}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)p_\theta(x_0|x_1)}{q(x_T|x_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\Big] \tag{32}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log p_\theta(x_0|x_1) + \log \frac{p(x_T)}{q(x_T|x_0)} + \sum_{t=2}^{T} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\Big] \tag{33}$$

$$= \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log p_\theta(x_0|x_1)\Big] + \mathbb{E}_{q(x_{1:T}|x_0)}\Big[\log \frac{p(x_T)}{q(x_T|x_0)}\Big] + \sum_{t=2}^{T}\Big[\mathbb{E}_{q(x_{1:T}|x_0)} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\Big] \tag{34}$$

$$= \mathbb{E}_{q(x_1|x_0)}\Big[\log p_\theta(x_0|x_1)\Big] + \mathbb{E}_{q(x_T|x_0)}\Big[\log \frac{p(x_T)}{q(x_T|x_0)}\Big] + \sum_{t=2}^{T}\Big[\mathbb{E}_{q(x_t,x_{t-1}|x_0)} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\Big] \tag{35}$$

$$= \underbrace{-KL(q(x_T|x_0)||p(x_T))}_{L_T} - \underbrace{\sum_{t=2}^{T}\mathbb{E}_{q(x_t|x_0)}\Big[KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))\Big]}_{L_{t-1}} + \underbrace{\mathbb{E}_{q(x_1|x_0)}\Big[\log p_\theta(x_0|x_1)\Big]}_{L_0} \tag{36}$$

To this end, we get the same formula as the Equation (5) in the DDPM paper, except for the negative sign. In the DDPM paper, the authors put a negative sign before the ELBO equation.

Similarly, L_T is just the difference between 2 Gaussian distributions, and thus do not require training. L_{t-1} is the difference between the diffusion step and the reverse step for all time steps, but this time, we have the grounth-truth image x_0 as input. L_0 is the same as that in Equation 22.

During the training process, we need to gradually add Gaussian noise to time step t for every training step. You can imagine that this process is time consuming. There is one key advantage of using Gaussian noise instead of other noise distributions. For an arbitrary time step t, instead of gradually adding noise, we can directly obtain the distribution at time step t by:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{(\bar{\alpha_t})}x_0, (1 - \bar{\alpha_t})\mathbf{I}) \tag{37}$$

Here we will demonstrate this. For an arbitrary time step t with added Gaussian noise ϵ:

$$\begin{aligned} x_t &= \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon'_{t-1} && (38)\\ &= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{1-\alpha_{t-1}}\epsilon'_{t-2}) + \sqrt{1-\alpha_t}\epsilon'_{t-1} && (39)\\ &= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\epsilon'_{t-2} + \sqrt{1-\alpha_t}\epsilon'_{t-1} && (40) \end{aligned}$$

Note that the sum of 2 normally distributed random variables is normal, with its mean being the sum of the two means, and its variance being the sum of the two variances (i.e., the square of the standard deviation is the sum of the squares of the standard deviations). With this property, we can further derive Equation 40.

$$= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\epsilon'_{t-2} + \sqrt{1-\alpha_t}\epsilon'_{t-1} \tag{41}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\epsilon_{t-2} \tag{42}$$

$$\text{the noise } \epsilon \text{ is different here, so we remove the } ' \tag{43}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1} + 1 - \alpha_t}\epsilon_{t-2} \tag{44}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{1 - \alpha_t\alpha_{t-1}}\epsilon_{t-2} \tag{45}$$

$$\text{Keep substituting } x_{t-2},\ x_{t-3} \ldots \tag{46}$$

$$= \sqrt{\prod_{s=1}^{t}\alpha_s}x_0 + \sqrt{1 - \prod_{s=1}^{t}\alpha_s}\epsilon_0 \tag{47}$$

$$\text{Remember the authors re-parameterize } \alpha_t := 1 - \beta_t \text{ and } \bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s \tag{48}$$

$$= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0 \tag{49}$$

$$\sim \mathcal{N}(x_t; \sqrt{(\bar{\alpha}_t)}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{50}$$

$$\text{Then we get Equation 37} \tag{51}$$

This means that to get the noisy image x_t at arbitrary time step t, instead of gradually adding noise, we can directly compute it, which save a lot of time during training.

Now, let's go back to the DDPM paper. In Equation (5) of the DDPM paper (or Equation 36 in this tutorial), it says this equation uses KL divergence to directly compare $p_\theta(x_{t-1}|x_t))$ against the forward diffusion process, which is $q(x_{t-1}|x_t, x_0)$. It says this is tractable when conditioned on x_0. This is easy to understand, if we know x_0, we definitely know x_{t-1} and x_t because both of them can be obtained just by adding Gaussian noise to x_0. In the DDPM paper, it says:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbf{I}) \tag{52}$$

$$\text{where } \tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \text{ and } \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{53}$$

Let's see how we get this. First, expanding $q(x_{t-1}|x_t, x_0)$ using Bayes rule.

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{54}$$

$$= \frac{\mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})\mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0), (1-\bar{\alpha}_{t-1})\mathbf{I}}{\mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})} \tag{55}$$

$$\text{Just expressing everything in its Gaussian form.} \tag{56}$$

$$\text{Note that the probability density function of Gaussian is: } \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{57}$$

$$\text{Plug it in and let's ignore the } \frac{1}{\sigma\sqrt{2\pi}} \text{ part} \tag{58}$$

$$\propto e^{-\frac{1}{2}(\frac{x_t - \sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}})^2} e^{-\frac{1}{2}(\frac{x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0}{\sqrt{1-\bar{\alpha}_{t-1}}})^2} \Big/ e^{-\frac{1}{2}(\frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}})^2} \tag{59}$$

$$= \exp\Big[-\frac{1}{2}(\frac{x_t - \sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}})^2 - \frac{1}{2}(\frac{x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0}{\sqrt{1-\bar{\alpha}_{t-1}}})^2 + \frac{1}{2}(\frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}})^2\Big] \tag{60}$$

$$= \exp\Big[-\frac{1}{2}\Big(\frac{(x_t - \sqrt{\alpha_t}x_{t-1})^2}{1-\alpha_t} + \frac{(x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(x_t - \sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t}\Big)\Big] \tag{61}$$

$$\text{Remember } (x-y)^2 = x^2 - 2xy + y^2\text{, use it to expend numerators} \tag{62}$$

$$\begin{aligned} &= \exp\Big\{-\frac{1}{2}\Big[\frac{(x_t^2 - 2\sqrt{\alpha_t}x_{t-1}x_t + \alpha_t x_{t-1}^2)}{1-\alpha_t} + \frac{(x_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}x_{t-1}x_0 + \bar{\alpha}_{t-1})}{1-\bar{\alpha}_{t-1}} \\ &- \frac{(x_t^2 - 2\sqrt{\bar{\alpha}_t}x_0x_t + \bar{\alpha}_t x_0^2)}{1-\bar{\alpha}_t}\Big]\Big\} \end{aligned} \tag{63}$$

$$\text{Let's ignore the terms that only depend on } x_0,\ x_t,\ \text{and different } \alpha \tag{64}$$

$$\text{As these are given terms with respect to } x_{t-1} \tag{65}$$

$$\propto \exp\Big\{-\frac{1}{2}\Big[\frac{(-2\sqrt{\alpha_t}x_{t-1}x_t + \alpha_t x_{t-1}^2)}{1-\alpha_t} + \frac{(x_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}x_{t-1}x_0)}{1-\bar{\alpha}_{t-1}}\Big]\Big\} \tag{66}$$

$$= \exp\Big\{-\frac{1}{2}\Big[\frac{-2\sqrt{\alpha_t}x_{t-1}x_t}{1-\alpha_t} + \frac{\alpha_t x_{t-1}^2}{1-\alpha_t} + \frac{x_{t-1}^2}{1-\bar{\alpha}_{t-1}} - \frac{2\sqrt{\bar{\alpha}_{t-1}}x_{t-1}x_0}{1-\bar{\alpha}_{t-1}}\Big]\Big\} \tag{67}$$

$$= \exp\Big\{-\frac{1}{2}\Big[x_{t-1}^2(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}) - 2x_{t-1}(\frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}} + \frac{\sqrt{\alpha_t}x_t}{1-\alpha_t})\Big]\Big\} \tag{68}$$

$$= \exp\Big\{-\frac{1}{2}\Big[x_{t-1}^2(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}) - 2x_{t-1}(\frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}} + \frac{\sqrt{\alpha_t}x_t}{1-\alpha_t})\Big]\Big\} \tag{69}$$

$$= \exp\Big\{-\frac{1}{2}\Big(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big)\Big[x_{t-1}^2 - 2x_{t-1}\frac{(\frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}} + \frac{\sqrt{\alpha_t}x_t}{1-\alpha_t})}{\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}}\Big]\Big\} \tag{70}$$

$$= \exp\Big\{-\frac{1}{2}\Big(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big)\Big[x_{t-1}^2 - 2x_{t-1}\frac{(\frac{\sqrt{\bar{\alpha}_{t-1}}x_0}{1-\bar{\alpha}_{t-1}} + \frac{\sqrt{\alpha_t}x_t}{1-\alpha_t})(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\Big]\Big\} \tag{71}$$

$$= \exp\Big\{-\frac{1}{2}\Big(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\Big)\Big[x_{t-1}^2 - 2x_{t-1}\frac{\sqrt{\bar{\alpha}_{t-1}}x_0(1-\alpha_t) + \sqrt{\alpha_t}x_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\Big]\Big\} \tag{72}$$

$$= \exp\Big\{-\frac{1}{2}\Big(\frac{1}{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\Big)\Big[x_{t-1}^2 - 2x_{t-1}\Big(\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t\Big)\Big]\Big\} \tag{73}$$

$$\text{Does it look familiar? Take a look at the PDF of Gaussian and Equation 53} \tag{74}$$

$$\propto \mathcal{N}\Big(x_{t-1}; \underbrace{\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t}_{\tilde{\mu}_t(x_t, x_0)}, \underbrace{\frac{(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}(1-\alpha_t)}_{\tilde{\beta}_t}\Big) \tag{75}$$

$$\text{Remember } 1-\alpha_t = \beta_t \tag{76}$$

To this end, we show how the Equation (7) in the DDPM paper was derived. The goal of training the diffusion model is to make $q(x_{t-1}|x_t, x_0)$ and $p_\theta(x_{t-1}|x_t)$ as close as possible. As discussed previously, we can express both as Gaussian. The KL divergence between 2 Gaussian distributions can be calculated as:

$$KL(\mathcal{N}_0||\mathcal{N}_1) = \frac{1}{2}\Big(\text{tr}(\sum_1^{-1}\sum_0) - k + (\mu_1 - \mu_0)^T \sum_1^{-1}(\mu_1 - \mu_0) + \ln(\frac{\det \sum_1}{\det \sum_0})\Big) \tag{77}$$

where μ, $\sum$, and k are means, covariance matrices, and dimensions of the Gaussian distributions.

In the DDPM paper, it says the variances can be learned or held as constant hyperparameters. Let's make it constant first, and set it to match the variances of the forward diffusion process. The goal of training diffusion model is to make the forward diffusion and backward denoising process as close as possible. Remember, the covariance of a random variable with itself is just the variance.

$$\arg \underset{\theta}{min}\ KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \tag{78}$$

$$= \arg \underset{\theta}{min}\ KL(\mathcal{N}(x_{t-1}; \tilde{\mu}_t, \sum_t)||\mathcal{N}(x_{t-1}; \tilde{\mu}_\theta, \sum_t)) \tag{79}$$

$$= \arg \underset{\theta}{min}\ \frac{1}{2}\Big[\ln 1 - k + \text{tr}(\sum_t^{-1}\sum_t) + (\tilde{\mu}_\theta - \tilde{\mu}_t)^T \sum_t^{-1}(\tilde{\mu}_\theta - \tilde{\mu}_t)\Big] \tag{80}$$

$$\text{Note that tr}(\sum_t^{-1}\sum_t) = k \tag{81}$$

$$= \arg \underset{\theta}{min}\ \frac{1}{2}\Big[(\tilde{\mu}_\theta - \tilde{\mu}_t)^T \sum_t^{-1}(\tilde{\mu}_\theta - \tilde{\mu}_t)\Big] \tag{82}$$

$$= \arg \underset{\theta}{min}\ \frac{1}{2}\Big[(\tilde{\mu}_\theta - \tilde{\mu}_t)^T(\sigma_t^2 \mathbf{I})^{-1}(\tilde{\mu}_\theta - \tilde{\mu}_t)\Big] \tag{83}$$

$$= \arg \underset{\theta}{min}\ \frac{1}{2\sigma_t^2}\Big[(\tilde{\mu}_\theta - \tilde{\mu}_t)^T \mathbf{I}(\tilde{\mu}_\theta - \tilde{\mu}_t)\Big] \tag{84}$$

$$\text{Only the diagonal part is kept, which is just L2} \tag{85}$$

$$= \arg \underset{\theta}{min}\ \frac{1}{2\sigma_t^2}\Big[||\tilde{\mu}_t - \tilde{\mu}_\theta||_2^2\Big] \tag{86}$$

Here, we reach the Equation (8) in the DDPM paper. We show that to train a diffusion model, we can just train the network so that the mean of the reverse denoising process is close to the mean of the forward process.

Remember that $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0$. And $x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon)$. And we also have $\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$. Let's plug them into Equation 86.

$$\arg \underset{\theta}{min}\ \frac{1}{2\sigma_t^2}\Big[||\tilde{\mu}_t(x_t, x_0) - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{87}$$

$$= \frac{1}{2\sigma_t^2}\Big[||\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{88}$$

$$= \frac{1}{2\sigma_t^2}\Big[||\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{89}$$

$$\text{Remember } \bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s \tag{90}$$

$$= \frac{1}{2\sigma_t^2}\Big[||\frac{\beta_t}{1-\bar{\alpha}_t}\frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{91}$$

$$= \frac{1}{2\sigma_t^2}\Big[||\frac{\beta_t\frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon) + \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t} - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{92}$$

$$= \frac{1}{2\sigma_t^2}\Big[||\frac{\beta_t(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon)}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t} - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{93}$$

$$= \frac{1}{2\sigma_t^2}\Big[||\frac{\beta_t x_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)} - \frac{\beta_t\sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t}{1-\bar{\alpha}_t} - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{94}$$

$$= \frac{1}{2\sigma_t^2}\Big[||\frac{\beta_t x_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)} - \frac{\beta_t\sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)} + \frac{\alpha_t(1-\bar{\alpha}_{t-1})x_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)} - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{95}$$

$$= \frac{1}{2\sigma_t^2}\Big[||x_t\frac{\beta_t + \alpha_t - \alpha_t\bar{\alpha}_{t-1}}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)} - \frac{\beta_t\sqrt{1-\bar{\alpha}_t}}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\epsilon - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{96}$$

$$\text{Note that } \alpha_t\bar{\alpha}_{t-1} = \bar{\alpha}_t \text{ and } \beta_t = 1 - \alpha_t \tag{97}$$

$$= \frac{1}{2\sigma_t^2}\Big[||x_t\frac{1-\bar{\alpha}_t}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)} - \frac{\beta_t\sqrt{1-\bar{\alpha}_t}}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\epsilon - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{98}$$

$$= \frac{1}{2\sigma_t^2}\Big[||x_t\frac{1}{\sqrt{\alpha_t}} - \frac{\beta_t\sqrt{1-\bar{\alpha}_t}}{\sqrt{\alpha_t}(1-\bar{\alpha}_t)}\epsilon - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{99}$$

$$= \frac{1}{2\sigma_t^2}\Big[||x_t\frac{1}{\sqrt{\alpha_t}} - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{100}$$

$$= \frac{1}{2\sigma_t^2}\Big[||\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon) - \tilde{\mu}_\theta(x_t, t)||_2^2\Big] \tag{101}$$

From Equation 101, we can see $\tilde{\mu}_\theta$ must predict $\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon)$ given x_t, which is consistent with the Equation (11) in the DDPM paper. With the predicted mean, and the fixed variance (can be learned or held as constant hyper-parameter). After training, the sampling process becomes:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon) + \sigma_t\mathbf{z} \tag{102}$$

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

You may noticed that in Algorithm 1 of the DDPM paper, the training is achieved by computing the difference between 2 noise variables. We can just substitute $\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon)$ into Equation 86 and train the neural network output $\epsilon(x_t, t)$ to be noise vector ϵ

$$\arg \underset{\theta}{min}\ \frac{1}{2\sigma_t^2}\Big[||\tilde{\mu}_t - \tilde{\mu}_\theta||_2^2\Big] \tag{103}$$

$$= \arg \underset{\theta}{min}\ \frac{1}{2\sigma_t^2}\Big[||\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon) - \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon(x_t, t))||_2^2\Big] \tag{104}$$

$$= \arg \underset{\theta}{min}\ \frac{1}{2\sigma_t^2}\Big[|| - \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon + \frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\epsilon(x_t, t))||_2^2\Big] \tag{105}$$

$$= \arg \underset{\theta}{min}\ \frac{1}{2\sigma_t^2}\frac{\beta_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\Big[||\epsilon(x_t, t) - \epsilon)||_2^2\Big] \tag{106}$$

$$= \arg \underset{\theta}{min} \frac{1}{2\sigma_t^2} \frac{\beta_t^2}{\alpha_t(1-\bar{\alpha}_t)} \left[||\epsilon(x_t, t) - \epsilon||_2^2 \right] \tag{107}$$

To this end, we obtained the Equation (12) in the DDPM paper. Equations 107 and 86 are the same objective. But as discussed in the DDPM paper, training using Equation 107 usually lead to better results. The objective of diffusion models is always Equation 78. If we make the model to output the noise ϵ, the mean at step t becomes $\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon(x_t, t))$. We can also make it to produce x_0 with some formulation changes, and the mean will become $\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$. Both ways are the same. In the DDPM paper, the network outputs noise, so their sampling is $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon(x_t, t)) + \sigma_t\mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$

This concludes this introductory tutorial for diffusion models. Hope it helps.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models." [Online]. Available: http://arxiv.org/abs/2006.11239

[2] C. Luo, "Understanding diffusion models: A unified perspective." [Online]. Available: http://arxiv.org/abs/2208.11970