Notes on Diffusion Models

Linghai Liu

August 31, 2023

Contents

1 Variational Autoencoder (VAE)

In Kingma and Welling [3], the authors divide the model into an encoder-decoder structure. Given a data point $\mathbf{x}$, the encoder estimates the distribution $q_\phi(\mathbf{z}\,|\,\mathbf{x})$ over the latent variable $\mathbf{z}$, while the decoder estimates the distribution over $\mathbf{x}$ given the latent variable $\mathbf{z}$. ϕ and θ are the parameters of the encoder and decoder, respectively.

Both the encoder and the decoder are parametrized by Multilayer Perceptrons (MLP), with a hidden dimension of 500. `ReLU` is used as activation, except for the last mapping from hidden state to reconstruction, where the `sigmoid` function is used to get pixel intensities in $[0, 1]$.

VAEs are trained by maximizing the log probability, $\log p(\mathbf{x})$:

$$\begin{aligned}
\log p(\mathbf{x}) &= \log p(\mathbf{x}) \int q_\phi(\mathbf{z}\,|\,\mathbf{x}) d\,\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[\log\left(\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{z}\,|\,\mathbf{x})}\cdot\frac{q_\phi(\mathbf{z}\,|\,\mathbf{x})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\right)\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[\log\frac{p(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\right] + D_{KL}(q_\phi(\mathbf{z}\,|\,\mathbf{x})\,||\,p(\mathbf{z}\,|\,\mathbf{x}))
\end{aligned} \tag{1.1}$$

The first term is called *evidence lower bound* (ELBO). Since the second term (usually referred to as *Kullback-Leibler divergence*, or *relative entropy*) is always nonnegative, maximizing the ELBO improves the lower bound of $\log p(x)$. Further decomposing ELBO by conditional probability, we have:

$$\text{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\left[\log\frac{p_\theta(\mathbf{x}\,|\,\mathbf{z})p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}\,|\,\mathbf{x})}\right] = -D_{KL}(q_\phi(\mathbf{z}\,|\,\mathbf{x})\,||\,p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{z}\,|\,\mathbf{x})}[\log p_\theta(\mathbf{x}\,|\,\mathbf{z})] \tag{1.2}$$

When modeling the latent distribution, we assume that it is Gaussian with diagonal covariance matrix, so we denote the estimated mean and main diagonal as $\boldsymbol{\mu}(\mathbf{x};\phi)$ and $\sigma^2(\mathbf{x};\phi)$ given a data point $\mathbf{x}$. By assuming that $\mathbf{z}$ encoded by $q_\phi(\cdot\,|\,\mathbf{x})$ should follow a standard J-dimensional multivariate Gaussian distribution, the KL-divergence term could be calculated analytically:

$$-D_{KL}(q_\phi(\mathbf{z}\,|\,\mathbf{x})\,||\,p_\theta(\mathbf{z})) = \frac{1}{2}\sum_{j=1}^{J}\big(1+\log\sigma_j^2(\mathbf{x};\phi)-\mu_j^2(\mathbf{x};\phi)+\sigma_j^2(\mathbf{x};\phi)\big) \tag{1.3}$$

and the other term from ELBO could be estimated via Monte Carlo. The authors of the original paper claim that a sample size of 1 is sufficient [3].

2 Diffusion Models

2.1 Denoising Diffusion Probabilistic Models (DDPMs)

Similar to VAEs, the diffusion model also consists of two processes - we refer to them as the forward process and the reverse process. The main differences are that the number of layers, T, is larger, and each latent variable has the same dimension as the input data $\mathbf{x}_0$.

In the forward process (denoted by q), we corrupt the input $\mathbf{x}\in\mathbb{R}^d$ (we suppose that it is an image) with noise via a noise-level schedule $\{\beta_t\}_{t=1}^{T}$, where $0<\beta_1<\ldots<\beta_T<1$. Each transition is both Gaussian and Markovian: let $\alpha_t = 1-\beta_t$. Then $\mathbf{x}_t \sim \mathcal{N}(\sqrt{\alpha_t}\,\mathbf{x}_{t-1},(1-\alpha_t)\mathbf{I})$. The goal of the forward process is to make $\mathbf{x}_T$ into unrecognizable noise, e.g., standard multivariate Gaussian.

In the reverse process (denoted by p_θ), we use a deep learning model parametrized by θ that reconstructs the original input by learning to estimate noise levels at each transition.

The joint distribution of the states $\mathbf{x}_0,\mathbf{x}_1,\ldots,\mathbf{x}_T$ during the forward and reverse process are

$$q(\mathbf{x}_0,\mathbf{x}_1,\ldots\mathbf{x}_T) = q(\mathbf{x}_0)\prod_{t=1}^{T} q(\mathbf{x}_t\,|\,\mathbf{x}_{t-1}), \quad p_\theta(\mathbf{x}_0,\mathbf{x}_1,\ldots,\mathbf{x}_T) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}\,|\,\mathbf{x}_t) \tag{2.1}$$

respectively. Note that by Bayes rule, we have

$$\begin{aligned} q(\mathbf{x}_t\,|\,\mathbf{x}_{t-1}) &= q(\mathbf{x}_t\,|\,\mathbf{x}_{t-1},\mathbf{x}_0) \quad \text{by Markov property} \\ &= \frac{q(\mathbf{x}_{t-1}\,|\,\mathbf{x}_t,\mathbf{x}_0)q(\mathbf{x}_t\,|\,\mathbf{x}_0)}{q(\mathbf{x}_{t-1}\,|\,\mathbf{x}_0)} \quad \text{by Bayes rule} \end{aligned} \tag{2.2}$$

The optimizing objective is, again, to maximize the expectation of log density $\log p_\theta(\mathbf{x}_0)$ over $q(\mathbf{x}_0,\mathbf{x}_1,\ldots,\mathbf{x}_T)$, and this objective can be decomposed as follows:

$$
\begin{aligned}
\log p_\theta(\mathbf{x}_0) &= \log \int \frac{p_\theta(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T) q(\mathbf{x}_1, \ldots, \mathbf{x}_T \,|\, \mathbf{x}_0)}{q(\mathbf{x}_1, \ldots, \mathbf{x}_T \,|\, \mathbf{x}_0)} d\,\mathbf{x}_1 \cdots d\,\mathbf{x}_T \\
&= \log \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\frac{p_\theta(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T)}{q(\mathbf{x}_1, \ldots, \mathbf{x}_T \,|\, \mathbf{x}_0)} \right] \\
&\geq \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T)}{q(\mathbf{x}_1, \ldots, \mathbf{x}_T \,|\, \mathbf{x}_0)} \right] \qquad \text{by Jensen's Inequality} \\
&= \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log \left(p(\mathbf{x}_T) \prod_{t=1}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t)}{q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})} \right) \right] \qquad \text{use Eq.2.1} \\
&= \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log \left(\frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1)}{q(\mathbf{x}_1 \,|\, \mathbf{x}_0)} \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t)}{q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})} \right) \right] \\
&= \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log \left(\frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1)}{q(\mathbf{x}_1 \,|\, \mathbf{x}_0)} \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t) q(\mathbf{x}_{t-1} \,|\, \mathbf{x}_0)}{q(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t \,|\, \mathbf{x}_0)} \right) \right] \qquad \text{use Eq.2.2} \\
&= \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log \left(\frac{p(\mathbf{x}_T) p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1) \cancel{q(\mathbf{x}_1 \,|\, \mathbf{x}_0)}}{\cancel{q(\mathbf{x}_1 \,|\, \mathbf{x}_0)} q(\mathbf{x}_T \,|\, \mathbf{x}_0)} \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t, \mathbf{x}_0)} \right) \right] \\
&= \underbrace{\mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1) \right]}_{C_0} + \underbrace{\mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T \,|\, \mathbf{x}_0)} \right]}_{C_T} \\
&\qquad + \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log \frac{p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1})}{q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)} \right]}_{C_t}
\end{aligned}
\tag{2.3}
$$

In Eq.2.3, we have three terms for further decomposition:

- Term C_0:

$$
\begin{aligned}
C_0 &= \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1) \right] \\
&= \int q(\mathbf{x}_1, \ldots, \mathbf{x}_T \,|\, \mathbf{x}_0) \log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1) d\,\mathbf{x}_1 \cdots d\,\mathbf{x}_T \\
&= \int q(\mathbf{x}_1 \,|\, \mathbf{x}_0) \log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1) d\,\mathbf{x}_1 \\
&= \mathbb{E}_{q(\mathbf{x}_1 \,|\, \mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1) \right]
\end{aligned}
$$

- Term C_T:

$$
\begin{aligned}
C_T &= \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T \,|\, \mathbf{x}_0)} \right] \\
&= \int q(\mathbf{x}_1, \ldots, \mathbf{x}_T \,|\, \mathbf{x}_0) \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T \,|\, \mathbf{x}_0)} d\,\mathbf{x}_1 \cdots d\,\mathbf{x}_T \\
&= \int q(\mathbf{x}_T \,|\, \mathbf{x}_0) \log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T \,|\, \mathbf{x}_0)} d\,\mathbf{x}_T \\
&= -D_{KL}(q(\mathbf{x}_T \,|\, \mathbf{x}_0) || p(\mathbf{x}_T))
\end{aligned}
$$

- Terms C_t for $t = 1, \ldots, T-1$:

$$\begin{aligned}
C_t &= \mathbb{E}_{q(\mathbf{x}_1,\ldots,\mathbf{x}_T \,|\, \mathbf{x}_0)}\left[\log \frac{p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1})}{q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)}\right] \\
&= \int q(\mathbf{x}_1, \ldots, \mathbf{x}_T \,|\, \mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1})}{q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)} d\mathbf{x}_1 \cdots d\mathbf{x}_T \\
&= \int q(\mathbf{x}_t, \mathbf{x}_{t+1} \,|\, \mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1})}{q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)} d\mathbf{x}_t \, d\mathbf{x}_{t+1} \\
&= \int q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0) \left[\int q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1})}{q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)} d\mathbf{x}_t\right] d\mathbf{x}_{t+1} \quad \text{by conditional probability} \\
&= \int q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0)(-D_{KL}(q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)||p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}))) d\mathbf{x}_{t+1} \\
&= -\mathbb{E}_{q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0)}[D_{KL}(q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)||p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}))]
\end{aligned}$$

Thus, the *variational lower bound* reads:

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_{q(\mathbf{x}_1 \,|\, \mathbf{x}_0)} [\log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1)] - D_{KL}(q(\mathbf{x}_T \,|\, \mathbf{x}_0)||p(\mathbf{x}_T)) - \sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0)}[D_{KL}(q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)||p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}))] \tag{2.4}$$

The above derivation is similar to those derived in other literature [2, 6, 4], and the three terms are sometimes interpreted as the *reconstruction term*, *prior matching term*, and *consistency terms* [4]. The consistency terms aims to match the corresponding steps in the forward process and the reverse process.

Now, we investigate the distributions $q(\mathbf{x}_t \,|\, \mathbf{x}_0)$ and posteriors $q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)$ for appropriate choices of t. Given $\mathbf{x}_0$ and t, we apply the forward process outlined by $\mathbf{x}_t \sim \mathcal{N}(\sqrt{\alpha_t}\,\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I})$. By the reparametrization trick [1] , we can sample an independent noise term $\varepsilon_t \sim \mathcal{N}(\mathbf{o}, \mathbf{I})$ and write

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

Recursively doing this:

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\,\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t \\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\,\mathbf{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\varepsilon_{t-1}) + \sqrt{1-\alpha_t}\varepsilon_t \\
&= \sqrt{\alpha_t\alpha_{t-1}}\,\mathbf{x}_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\varepsilon_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t \\
&= \sqrt{\alpha_t\alpha_{t-1}}\,\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\tilde{\varepsilon}_2 \quad \text{by combining two independent noise (terms in cyan), } \tilde{\varepsilon}_2 \sim \mathcal{N}(\mathbf{o}, \mathbf{I}) \\
&= \ \vdots \\
&= \sqrt{\prod_{s=1}^{t} \alpha_s}\,\mathbf{x}_0 + \sqrt{1 - \prod_{s=1}^{t} \alpha_t}\tilde{\varepsilon}_t \\
&= \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\tilde{\varepsilon}_t
\end{aligned}$$

by letting $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Hence, we conclude that $q(\mathbf{x}_t \,|\, \mathbf{x}_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$.

[1] Reparametrization Trick: Suppose we have $\varepsilon \sim \mathcal{N}(\mathbf{o}, \mathbf{I})$, then $\mathbf{x} \overset{\mathcal{D}}{=} \boldsymbol{\mu} + \sigma\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. This is also useful during training, since gradients can go through the learned $\boldsymbol{\mu}_\theta$ and σ_θ.

Then, by Bayes rule, we compute $q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t \,|\, \mathbf{x}_0)}{q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0)}$ relying on the Markovian assumption of Gaussian forward transitions and the calculation of $q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0)$ above.

$$\begin{aligned}
q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0) &= \frac{q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t \,|\, \mathbf{x}_0)}{q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0)} \\
&\propto \exp\left\{ -\frac{1}{2} \left[\frac{||\, \mathbf{x}_{t+1} - \sqrt{\alpha_{t+1}}\, \mathbf{x}_t \,||^2}{1 - \alpha_{t+1}} + \frac{||\, \mathbf{x}_t - \sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 \,||^2}{1 - \bar{\alpha}_t} - \frac{||\, \mathbf{x}_{t+1} - \sqrt{\bar{\alpha}_{t+1}}\, \mathbf{x}_0 \,||^2}{1 - \bar{\alpha}_{t+1}} \right] \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \left[\left(\frac{\alpha_{t+1}}{1 - \alpha_{t+1}} + \frac{1}{1 - \bar{\alpha}_t} \right) ||\, \mathbf{x}_t \,||^2 + \left(\frac{-2\sqrt{\alpha_{t+1}}\, \mathbf{x}_{t+1}}{1 - \alpha_{t+1}} + \frac{-2\sqrt{\bar{\alpha}_t}\, \mathbf{x}_0}{1 - \bar{\alpha}_t} \right)^T \mathbf{x}_t \right] \right\} \\
&= \exp\left\{ -\frac{1}{2} \left[\frac{1 - \bar{\alpha}_{t+1}}{(1 - \alpha_{t+1})(1 - \bar{\alpha}_t)} ||\, \mathbf{x}_t \,||^2 - 2 \left(\frac{\sqrt{\alpha_{t+1}}(1 - \bar{\alpha}_t)\, \mathbf{x}_{t+1} + \sqrt{\bar{\alpha}_t}(1 - \alpha_{t+1})\, \mathbf{x}_0}{(1 - \alpha_{t+1})(1 - \bar{\alpha}_t)} \right)^T \mathbf{x}_t \right] \right\}
\end{aligned}$$

The density of a multivariate Gaussian distribution with covariance like $\sigma^2 \mathbf{I}$ is proportional to

$$\exp\left(-\frac{1}{2\sigma^2} ||\, \mathbf{x} - \boldsymbol{\mu} ||^2 \right)$$

so $q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0) \sim \mathcal{N}\left(\frac{\sqrt{\alpha_{t+1}}(1 - \bar{\alpha}_t)\, \mathbf{x}_{t+1} + \sqrt{\bar{\alpha}_t}(1 - \alpha_{t+1})\, \mathbf{x}_0}{1 - \bar{\alpha}_{t+1}}, \frac{(1 - \alpha_{t+1})(1 - \bar{\alpha}_t)}{1 - \bar{\alpha}_{t+1}} \mathbf{I} \right)$.

Before we further derive the training objective (Eq.2.4), we have to define p_θ. Since we know the noise schedule, it suffices to learn the mean vectors at different t if we assume the reverse process also consists of Gaussian transitions for convenience. We can have a neural network (parametrized by θ) that either estimates the mean vector directly or estimate the noise given $\mathbf{x}_t$ and t, i.e., $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ or $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$.

We first naïvely choose $\boldsymbol{\mu}_\theta(\mathbf{x}_{t+1}, t+1)$, i.e., $p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}) \sim \mathcal{N}\left(\boldsymbol{\mu}_\theta(\mathbf{x}_{t+1}, t+1), \frac{(1 - \alpha_{t+1})(1 - \bar{\alpha}_t)}{1 - \bar{\alpha}_{t+1}} \mathbf{I} \right)$. By Appendix A,

$$\begin{aligned}
&D_{KL}(q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0) || p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1})) \\
=&\frac{1}{2} \left(\log \frac{\det(\Sigma_2)}{\det(\Sigma_1)} - d + \mathrm{Tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) \right) \quad \text{with } \mu_1, \mu_2, \Sigma_1, \Sigma_2 \text{ to be plugged in.} \\
=&\frac{1 - \bar{\alpha}_{t+1}}{2(1 - \alpha_{t+1})(1 - \bar{\alpha}_t)} \left|\left| \boldsymbol{\mu}_\theta(\mathbf{x}_{t+1}, t+1) - \frac{\sqrt{\alpha_{t+1}}(1 - \bar{\alpha}_t)\, \mathbf{x}_{t+1} + \sqrt{\bar{\alpha}_t}(1 - \alpha_{t+1})\, \mathbf{x}_0}{1 - \bar{\alpha}_{t+1}} \right|\right|^2
\end{aligned} \tag{2.5}$$

One can also try to estimate the clean, original image $\mathbf{x}_0$ [4] starting from $(\mathbf{x}_t, t)$ with a network $\mathbf{x}_\theta(\mathbf{x}_t, t)$:

$$p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}) \sim \mathcal{N}\left(\frac{\sqrt{\alpha_{t+1}}(1 - \bar{\alpha}_t)\, \mathbf{x}_{t+1} + \sqrt{\bar{\alpha}_t}(1 - \alpha_{t+1})\, \mathbf{x}_\theta(\mathbf{x}_{t+1}, t+1)}{1 - \bar{\alpha}_{t+1}}, \frac{(1 - \alpha_{t+1})(1 - \bar{\alpha}_t)}{1 - \bar{\alpha}_{t+1}} \mathbf{I} \right)$$

Then Eq.2.5 becomes

$$\frac{\bar{\alpha}_t(1 - \alpha_{t+1})}{2(1 - \bar{\alpha}_t)(1 - \bar{\alpha}_{t+1})} \left|\left| \mathbf{x}_\theta(\mathbf{x}_{t+1}, t+1) - \mathbf{x}_0 \right|\right|^2 \tag{2.6}$$

Alternatively, we can adopt $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)$ [2]. We use the reparametrization trick for $q(\mathbf{x}_t \,|\, \mathbf{x}_0)$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\, \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \tilde{\boldsymbol{\varepsilon}}_t \Longleftrightarrow \mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \tilde{\boldsymbol{\varepsilon}}_t}{\sqrt{\bar{\alpha}_t}} \tag{2.7}$$

Replacing $\mathbf{x}_0$ in the posterior mean:

$$\begin{aligned}
&\frac{\sqrt{\alpha_{t+1}}(1 - \bar{\alpha}_t)\, \mathbf{x}_{t+1} + \sqrt{\bar{\alpha}_t}(1 - \alpha_{t+1})\, \mathbf{x}_0}{1 - \bar{\alpha}_{t+1}} \\
=&\frac{\sqrt{\alpha_{t+1}}(1 - \bar{\alpha}_t)\, \mathbf{x}_{t+1} + \sqrt{\bar{\alpha}_t}(1 - \alpha_{t+1}) \frac{\mathbf{x}_{t+1} - \sqrt{1 - \bar{\alpha}_{t+1}} \tilde{\boldsymbol{\varepsilon}}_{t+1}}{\sqrt{\bar{\alpha}_{t+1}}}}{1 - \bar{\alpha}_{t+1}} \\
=&\frac{1}{\sqrt{\alpha_{t+1}}} \mathbf{x}_{t+1} - \frac{1 - \alpha_{t+1}}{\sqrt{(1 - \bar{\alpha}_{t+1}) \alpha_{t+1}}} \tilde{\boldsymbol{\varepsilon}}_{t+1}
\end{aligned} \tag{2.8}$$

As $\varepsilon_\theta(\mathbf{x}_t, t)$ estimates $\tilde{\varepsilon}_t$, Eq.2.5 reads (by plugging in Eq.2.8):

$$\begin{aligned}
&D_{KL}(q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)||p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1})) \\
=&\frac{1-\bar{\alpha}_{t+1}}{2(1-\alpha_{t+1})(1-\bar{\alpha}_t)}\left|\left|\frac{1-\alpha_{t+1}}{\sqrt{(1-\bar{\alpha}_{t+1})\alpha_{t+1}}}(\varepsilon_\theta(\mathbf{x}_{t+1}, t+1)-\tilde{\varepsilon}_{t+1})\right|\right|^2 \\
=&\frac{\cancel{1-\bar{\alpha}_{t+1}}}{2\cancel{(1-\alpha_{t+1})}(1-\bar{\alpha}_t)}\cdot\frac{(1-\alpha_{t+1})^{\cancel{2}}}{\cancel{(1-\bar{\alpha}_{t+1})}\alpha_{t+1}}\left|\left|\varepsilon_\theta(\mathbf{x}_{t+1}, t+1)-\tilde{\varepsilon}_{t+1}\right|\right|^2 \\
=&\frac{1-\alpha_{t+1}}{2\alpha_{t+1}(1-\bar{\alpha}_t)}\left|\left|\varepsilon_\theta(\mathbf{x}_{t+1}, t+1)-\tilde{\varepsilon}_{t+1}\right|\right|^2
\end{aligned} \tag{2.9}$$

In Ho et al. [2], the authors assumed that the original image $\mathbf{x}_0$ being scaled within the interval $[-1, 1]$, and calculated $\log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1)$ by integrating the normal density in each of the d dimensions separately:

$$\begin{aligned}
&p(\mathbf{x}_0 \,|\, \mathbf{x}_1) = \prod_{i=1}^{d}\int_{\delta_-(\mathbf{x}_{0,i})}^{\delta_+(\mathbf{x}_{0,i})}\mathcal{N}(x_i; \boldsymbol{\mu}_\theta(\mathbf{x}_1, 1), \sigma_1^2)dx_i \\
&\delta_+(x) = \begin{cases}\infty & \text{if } x = 1 \\ x+\frac{1}{255} & \text{if } x < 1\end{cases} \qquad \delta_-(x) = \begin{cases}-\infty & \text{if } x = -1 \\ x+\frac{1}{255} & \text{if } x > -1\end{cases}
\end{aligned} \tag{2.10}$$

Note that $\boldsymbol{\mu}_\theta(\mathbf{x}_1, 1)$ learns to return a good estimate of $\mathbf{x}_0$ given $\mathbf{x}_1$ and time 1. Hence, if our network estimates noise levels at different time t as in Eq.2.9, then in Eq.2.10, given $\mathbf{x}_0$, $\boldsymbol{\mu}_\theta(\mathbf{x}_1, 1) = \mathbf{x}_1 - \varepsilon_\theta(\mathbf{x}_1, 1)$.

$$\begin{aligned}
\log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1) &= \sum_{i=1}^{d}\log\int_{\delta_-(\mathbf{x}_{0,i})}^{\delta_+(\mathbf{x}_{0,i})}\mathcal{N}(x_i; \boldsymbol{\mu}_\theta(\mathbf{x}_1, 1), \sigma_1^2)dx_i \\
&\approx -\sum_{i=1}^{d}\frac{2}{255}\cdot\frac{1}{2\sigma_1^2}(\underbrace{\mathbf{x}_{0,i}-\mathbf{x}_{1,i}}_{:=-\tilde{\varepsilon}_{1,i}}+\varepsilon_\theta(\mathbf{x}_1, 1)_i)^2 + C \quad \text{for some constant } C \\
&= -\gamma_0\left|\left|\varepsilon_\theta(\mathbf{x}_1, 1)-\tilde{\varepsilon}_1\right|\right|^2 + C \quad \text{for some constant } \gamma_0
\end{aligned} \tag{2.11}$$

Hence, in order to learn a "good" distribution $p_\theta(\mathbf{x}_0)$, we have to maximize the variational lower bound in Eq.2.4. By our calculations above, using $\varepsilon_\theta(\cdot, \cdot)$, our training objective becomes:

$$\begin{aligned}
&\max_\theta \mathbb{E}_{q(\mathbf{x}_1 \,|\, \mathbf{x}_0)}\left[\log p_\theta(\mathbf{x}_0 \,|\, \mathbf{x}_1)\right] - D_{KL}(q(\mathbf{x}_T \,|\, \mathbf{x}_0)||p(\mathbf{x}_T)) - \sum_{t=1}^{T-1}\mathbb{E}_{q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0)}\left[D_{KL}(q(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)||p_\theta(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}))\right] \\
\Longleftrightarrow &\max_\theta \mathbb{E}_{q(\mathbf{x}_1 \,|\, \mathbf{x}_0)}\left[-\gamma_0\left|\left|\varepsilon_\theta(\mathbf{x}_1, 1)-\tilde{\varepsilon}_1\right|\right|^2 + C\right] - \sum_{t=1}^{T-1}\mathbb{E}_{q(\mathbf{x}_{t+1} \,|\, \mathbf{x}_0)}\left[\frac{1-\alpha_{t+1}}{2\alpha_{t+1}(1-\bar{\alpha}_t)}\left|\left|\varepsilon_\theta(\mathbf{x}_{t+1}, t+1)-\tilde{\varepsilon}_{t+1}\right|\right|^2\right] \\
\Longleftrightarrow &\min_\theta \sum_{t=1}^{T}\mathbb{E}_{q(\mathbf{x}_t \,|\, \mathbf{x}_0)}\left[\gamma_t\left|\left|\varepsilon_\theta(\mathbf{x}_t, t)-\tilde{\varepsilon}_t\right|\right|^2\right] \quad \text{for some constants } \{\gamma_t\}_{t=1}^{T} \\
\Longleftrightarrow &\min_\theta L_\gamma(\theta)
\end{aligned} \tag{2.12}$$

2.2 Denoising Diffusion Implicit Models (DDIMs)

The authors of DDIM [5] observe that in Eq.2.12, what really matters when training a DDPM model is the marginal distribution $q(\mathbf{x}_t \,|\, \mathbf{x}_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$, rather than the joint distribution $q(\mathbf{x}_1, \ldots, \mathbf{x}_T \,|\, \mathbf{x}_0)$.

Also, in the DDPM paradigm, we hypothesize that the forward process indexed by $\{\alpha_t\}_{t=1}^T \subseteq (0,1)^T$ is Markovian, otherwise we cannot calculate the posterior distribution using Bayes theorem (as for DDPM) because $q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1})$ does not equal to $q(\mathbf{x}_t \,|\, \mathbf{x}_{t-1}, \mathbf{x}_0)$ in general.

The authors of DDIM used another vector $\{\sigma_t\}_{t=1}^T \subseteq [0,1)^T$ to define a family $\mathcal{Q}$ of distributions that satisfies:

$$q_\sigma(\mathbf{x}_1, \dots, \mathbf{x}_T \,|\, \mathbf{x}_0) = q_\sigma(\mathbf{x}_T \,|\, \mathbf{x}_0) \prod_{t=1}^{T-1} q_\sigma(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0) \tag{2.13}$$

while ensuring $q_\sigma(\mathbf{x}_t \,|\, \mathbf{x}_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$ for $t = 1, \dots, T$, as in DDPM. With this factorization, we would have the same variational lower bound as in Eq.2.4.

We have two goals now. The first goal is to find the posterior mean and covariance for $q_\sigma(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)$ in Eq.2.13 that gives the same $q(\mathbf{x}_t \,|\, \mathbf{x}_0)$ as in DDPM. The second goal is to derive a suitable training objective in the non-Markovian scheme starting from Eq.2.4.

In the DDIM paper [5], the authors proposed the following lemma:

Lemma 2.1 (Posterior for $\mathcal{Q}$)**.** Given $\{\alpha_t\}_{t=1}^T$ and $\{\sigma_t\}_{t=1}^T$, for all $q_\sigma \in \mathcal{Q}$, if

$$q_\sigma(\mathbf{x}_T \,|\, \mathbf{x}_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_T}\,\mathbf{x}_0, (1-\bar{\alpha}_T)\mathbf{I})$$

and for all $t = 1, \dots, T-1$,

$$q_\sigma(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0) \sim \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{\frac{1-\bar{\alpha}_t-\sigma_{t+1}^2}{1-\bar{\alpha}_{t+1}}}(\mathbf{x}_{t+1} - \sqrt{\bar{\alpha}_{t+1}}\,\mathbf{x}_0), \sigma_{t+1}^2\mathbf{I}\right)$$

then for all $t = 1, \dots, T$, we have

$$q_\sigma(\mathbf{x}_t \,|\, \mathbf{x}_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

Proof. The proof is done by induction from $t = T$ to $t = 1$. The base case is $t = T$, which is already given. Suppose the statement holds for $2 \le t = k \le T$. We want to show that the statement also holds for $t = k-1$. Denote $X = \mathbf{x}_k \,|\, \mathbf{x}_0$ and $Y = \mathbf{x}_{k-1} \,|\, \mathbf{x}_0$. By the definition of $q_\sigma(\mathbf{x}_t \,|\, \mathbf{x}_{t+1}, \mathbf{x}_0)$, $Y|X$ is Gaussian. Hence, we can use the result in Appendix B as follows:

$$\mu_X = \sqrt{\bar{\alpha}_k}\,\mathbf{x}_0, \quad \Sigma_X = (1-\bar{\alpha}_k)\mathbf{I}, \quad \Sigma_{Y|X} = \sigma_k^2\mathbf{I},$$
$$A = \sqrt{\frac{1-\bar{\alpha}_{k-1}-\sigma_k^2}{1-\bar{\alpha}_k}}\mathbf{I},$$
$$b = \sqrt{\bar{\alpha}_{k-1}}\,\mathbf{x}_0 - \sqrt{\frac{1-\bar{\alpha}_{k-1}-\sigma_k^2}{1-\bar{\alpha}_k}}\sqrt{\bar{\alpha}_k}\,\mathbf{x}_0$$

The marginal mean and covariance of Y are then given as:

$$\begin{aligned}
\mu_Y &= A\mu_X + b \\
&= \cancel{\sqrt{\frac{1-\bar{\alpha}_{k-1}-\sigma_k^2}{1-\bar{\alpha}_k}}}\sqrt{\bar{\alpha}_k}\,\mathbf{x}_0 + \sqrt{\bar{\alpha}_{k-1}}\,\mathbf{x}_0 - \cancel{\sqrt{\frac{1-\bar{\alpha}_{k-1}-\sigma_k^2}{1-\bar{\alpha}_k}}}\sqrt{\bar{\alpha}_k}\,\mathbf{x}_0 \\
&= \sqrt{\bar{\alpha}_{k-1}}\,\mathbf{x}_0 \\
\Sigma_Y &= \Sigma_{Y|X} + A\Sigma_X A^T \\
&= \cancel{\sigma_k^2}\mathbf{I} + \frac{1-\bar{\alpha}_{k-1}-\cancel{\sigma_k^2}}{\cancel{1-\bar{\alpha}_k}}\cancel{(1-\bar{\alpha}_k)}\mathbf{I} \\
&= (1-\bar{\alpha}_{k-1})\mathbf{I}
\end{aligned}$$

Therefore, $q_\sigma(\mathbf{x}_{k-1}\,|\,\mathbf{x}_k,\mathbf{x}_0) \sim \mathcal{N}(\sqrt{\bar\alpha_{k-1}}\,\mathbf{x}_0, (1-\bar\alpha_{k-1})\mathbf{I})$, and we finished the proof. □

The introduction of σ into our notation gives us a wider class of models that we can consider than just DDPMs, which features Markovian transitions. To make the transitions Markovian, we can just match the posterior covariance in both cases, which leads to

$$\sigma_{t+1}^2\mathbf{I} = \frac{(1-\alpha_{t+1})(1-\bar\alpha_t)}{1-\bar\alpha_{t+1}}\mathbf{I} \Longrightarrow \sigma_{t+1} = \sqrt{\frac{(1-\alpha_{t+1})(1-\bar\alpha_t)}{1-\bar\alpha_{t+1}}}, \quad t = 1,\dots,T-1 \tag{2.14}$$

For completeness, we can define an auxiliary parameter $\alpha_0 = 1$, so that $\sigma_1 = 0$.

When $\sigma_t \equiv 0$ for all t, the covariance of the posteriors $q(\mathbf{x}_t\,|\,\mathbf{x}_{t+1},\mathbf{x}_0)$ becomes **0**. This makes the forward process deterministic, in the sense that when we know both $\mathbf{x}_t$ and $\mathbf{x}_0$, we can solve for $\mathbf{x}_{t+1}$ by solving the condition in Lemma 2.1 (except for calculating or sampling $\mathbf{x}_1$):

$$\mathbf{x}_t = \sqrt{\bar\alpha_t}\,\mathbf{x}_0 + \sqrt{\frac{1-\bar\alpha_t}{1-\bar\alpha_{t+1}}}(\mathbf{x}_{t+1} - \sqrt{\bar\alpha_{t+1}}\,\mathbf{x}_0)$$

In this case (where $\sigma_t \equiv 0$ for all t), the model is called **denoising diffusion implicit model** (DDIM), and it is trained with the DDPM objective (Eq.2.12). The forward process in this case is no longer a diffusion: once $\mathbf{x}_1$ is sampled from $q_\alpha(\mathbf{x}_1\,|\,\mathbf{x}_0) \sim \mathcal{N}(\sqrt{\alpha_1}, (1-\alpha_1)\mathbf{I})$, $\mathbf{x}_2,\dots,\mathbf{x}_T$ are solved iteratively.

As in DDPM, we are trying to match $p_\theta(\mathbf{x}_t\,|\,\mathbf{x}_{t+1})$ and $q_\sigma(\mathbf{x}_t\,|\,\mathbf{x}_{t+1},\mathbf{x}_0)$. The covariance of the posterior is constant, so we only need to estimate the mean $\mu_\theta(\mathbf{x}_{t+1}, t+1)$ for $p_\theta(\mathbf{x}_t\,|\,\mathbf{x}_{t+1}), t = 1,\dots,T-1$.

The first step is same as that for DDPM: since $q(\mathbf{x}_{t+1}\,|\,\mathbf{x}_0) \sim \mathcal{N}(\sqrt{\bar\alpha_{t+1}}\,\mathbf{x}_0, (1-\bar\alpha_{t+1})\mathbf{I})$, we can sample $\tilde\varepsilon_{t+1} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and let $\mathbf{x}_{t+1} \overset{\mathcal{D}}{=} \sqrt{\bar\alpha_{t+1}}\,\mathbf{x}_0 + \sqrt{1-\bar\alpha_{t+1}}\tilde\varepsilon_{t+1} \Longrightarrow \mathbf{x}_0 = \frac{1}{\sqrt{\bar\alpha_{t+1}}}\mathbf{x}_{t+1} - \sqrt{\frac{1-\bar\alpha_{t+1}}{\bar\alpha_{t+1}}}\tilde\varepsilon_{t+1}$.

Plugging it into the posterior mean, we have:

$$\begin{aligned}
&\sqrt{\bar\alpha_t}\,\mathbf{x}_0 + \sqrt{\frac{1-\bar\alpha_t-\sigma_{t+1}^2}{1-\bar\alpha_{t+1}}}(\mathbf{x}_{t+1} - \sqrt{\bar\alpha_{t+1}}\,\mathbf{x}_0)\\
=&\sqrt{\frac{1-\bar\alpha_t-\sigma_{t+1}^2}{1-\bar\alpha_{t+1}}}\,\mathbf{x}_{t+1} + \left(\sqrt{\bar\alpha_t} - \sqrt{\frac{\bar\alpha_{t+1}(1-\bar\alpha_t-\sigma_{t+1}^2)}{1-\bar\alpha_{t+1}}}\right)\mathbf{x}_0\\
=&\sqrt{\frac{1-\bar\alpha_t-\sigma_{t+1}^2}{1-\bar\alpha_{t+1}}}\,\mathbf{x}_{t+1} + \left(\sqrt{\bar\alpha_t} - \sqrt{\frac{\bar\alpha_{t+1}(1-\bar\alpha_t-\sigma_{t+1}^2)}{1-\bar\alpha_{t+1}}}\right)\left(\frac{1}{\sqrt{\bar\alpha_{t+1}}}\mathbf{x}_{t+1} - \sqrt{\frac{1-\bar\alpha_{t+1}}{\bar\alpha_{t+1}}}\tilde\varepsilon_{t+1}\right)\\
=&\frac{1}{\sqrt{\alpha_{t+1}}}\mathbf{x}_{t+1} + \left(\sqrt{1-\bar\alpha_t-\sigma_{t+1}^2} - \sqrt{\frac{1-\bar\alpha_{t+1}}{\alpha_{t+1}}}\right)\tilde\varepsilon_{t+1}
\end{aligned}$$

We can use a network $\varepsilon_\theta(\mathbf{x}_{t+1}, t+1)$ to estimate $\tilde\varepsilon_{t+1}$, and let the mean and covariance of $p_\theta(\mathbf{x}_t\,|\,\mathbf{x}_{t+1})$ be $\mu_\theta(\mathbf{x}_{t+1}, t+1) = \frac{1}{\sqrt{\alpha_{t+1}}}\mathbf{x}_{t+1} + \left(\sqrt{1-\bar\alpha_t-\sigma_{t+1}^2} - \sqrt{\frac{1-\bar\alpha_{t+1}}{\alpha_{t+1}}}\right)\varepsilon_\theta(\mathbf{x}_{t+1}, t+1)$ and $\sigma_{t+1}^2\mathbf{I}$. Hence,

$$D_{KL}(q(\mathbf{x}_t\,|\,\mathbf{x}_{t+1},\mathbf{x}_0)||p_\theta(\mathbf{x}_{t+1}\,||\,\mathbf{x}_t)) = \left(\frac{\sqrt{1-\bar\alpha_t-\sigma_{t+1}^2} - \sqrt{\frac{1-\bar\alpha_{t+1}}{\alpha_{t+1}}}}{\sigma_{t+1}}\right)^2 \left|\left|\varepsilon_\theta(\mathbf{x}_{t+1}, t+1) - \tilde\varepsilon_{t+1}\right|\right|^2$$

Combining this with Eq.2.11, we can rewrite the training objective (Eq.2.4) as maximizing

$$\mathbb{E}_{q(\mathbf{x}_1\,|\,\mathbf{x}_0)}\left[-\gamma_0\left|\left|\varepsilon_\theta(\mathbf{x}_1, 1) - \tilde\varepsilon_1\right|\right|^2\right] - \sum_{t=1}^{T-1}\mathbb{E}_{q(\mathbf{x}_{t+1}\,|\,\mathbf{x}_0)}\left[\left(\frac{\sqrt{1-\bar\alpha_t-\sigma_{t+1}^2} - \sqrt{\frac{1-\bar\alpha_{t+1}}{\alpha_{t+1}}}}{\sigma_{t+1}}\right)^2\left|\left|\varepsilon_\theta(\mathbf{x}_{t+1}, t+1) - \tilde\varepsilon_{t+1}\right|\right|^2\right]$$

which is equivalent to minimizing

$$\sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t \mid \mathbf{x}_0)} \left[\gamma_t(\sigma) \left|\left| \varepsilon_\theta(\mathbf{x}_t, t) - \tilde{\varepsilon}_t \right|\right|^2 \right] := J_\sigma(\theta) \tag{2.15}$$

Note that Eq.2.15 has the same form as Eq.2.12 up to a constant (with respect to θ), so we can train DDIM models using the same objective as DDPM models [2].

Hence, our two goals are both achieved. Our next focus is on the generation process of DDIM. In the DDPM paper [2], the authors claimed to use $\gamma = \mathbf{1}$: $L_{\mathbf{1}}(\theta)$ can be used as a surrogate objective for $J_\sigma(\theta)$[5]. Hence, when our posterior is defined to maintain the same $q(\mathbf{x}_t \mid \mathbf{x}_0)$, we are able to consider forward processes shorter than T, i.e., we can choose a subset of times $\{1, \dots, T\}$, $\tau = \{\tau_1, \dots, \tau_S\} \subseteq \{1, \dots, T\}$, for some $1 \leq S < T$. In this manner, the sampling process becomes much more efficient than the original procedure, because in the original DDPM, both the forward process and the reverse process would take the full T steps.

For numerical experiments, the authors considered different subsequences τ and defined $\{\sigma_{\tau_i}\}_{i=1}^{S}$ as:

$$\sigma_{\tau_i} = \eta \sqrt{\frac{(1 - \alpha_{\tau_i})(1 - \bar{\alpha}_{\tau_i - 1})}{1 - \bar{\alpha}_{\tau_i}}}$$

with some $\eta \in [0, 1]$. According to condition 2.14, we get a DDIM model when $\eta = 0$; a DDPM model when $\eta = 1$; and a model with some stochasticity for values of η in between. The authors of DDIM [5] varied S, the length of the subsequence, and η, the parameter controlling stochasticity, generate samples from a trained DDIM model (trained with the DDPM objective), and then evaluated the model's quality based on Frechet Inception Distance (FID), along with a discussion on sample efficiency and consistency.

References

[1] Huixuan GAO. *Ying yong duo yuan tong ji fen xi (Applied Multivariate Statistical Analysis)*. PEKING UNIVERSITY PRESS, 2005.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[4] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[6] Lilian Weng. What are diffusion models? *lilianweng.github.io*, Jul 2021.

[2] This conclusion is the same idea as Theorem 1 in [5].

A KL Divergence between Gaussians

Let p and q be the probability density of $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, respectively. Then, $\forall x \in \mathbb{R}^d$,

$$\begin{aligned} p(x) &= \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\text{det}(\Sigma_1)}} \exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right\} \\ q(x) &= \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\text{det}(\Sigma_2)}} \exp\left\{-\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right\} \end{aligned}$$

The KL divergence becomes [3]:

$$\begin{aligned} D_{KL}(p||q) &= \int p(x)\log\frac{p(x)}{q(x)}dx \\ &= \int p(x)\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) + \log\sqrt{\frac{\text{det}(\Sigma_2)}{\text{det}(\Sigma_1)}}\right)dx \\ &= \frac{1}{2}\log\frac{\text{det}(\Sigma_2)}{\text{det}(\Sigma_1)} - \frac{1}{2}\mathbb{E}_{X\sim p}\left[\text{Tr}\left((X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)\right)\right] + \frac{1}{2}\mathbb{E}_{X\sim p}\left[\text{Tr}\left((X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)\right)\right] \end{aligned}$$

For the second term:

$$\begin{aligned} &-\frac{1}{2}\mathbb{E}_{X\sim p}\left[\text{Tr}\left((X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)\right)\right] \\ =& -\frac{1}{2}\mathbb{E}_{X\sim p}\left[\text{Tr}\left(\Sigma_1^{-1}(X-\mu_1)(X-\mu_1)^T\right)\right] \\ =& -\frac{1}{2}\,\text{Tr}\left(\Sigma_1^{-1}\mathbb{E}_{X\sim p}\left[(X-\mu_1)(X-\mu_1)^T\right]\right) \\ =& -\frac{1}{2}\,\text{Tr}(\Sigma_1^{-1}\Sigma_1) = -\frac{d}{2} \end{aligned}$$

For the third term:

$$\begin{aligned} &\frac{1}{2}\mathbb{E}_{X\sim p}\left[\text{Tr}\left((X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)\right)\right] \\ =&\frac{1}{2}\mathbb{E}_{X\sim p}\left[\text{Tr}\left(\Sigma_2^{-1}(X-\mu_2)(X-\mu_2)^T\right)\right] \\ =&\frac{1}{2}\mathbb{E}_{X\sim p}\left[\text{Tr}\left(\Sigma_2^{-1}((X-\mu_1)+(\mu_1-\mu_2))((X-\mu_1)+(\mu_1-\mu_2))^T\right)\right] \\ =&\frac{1}{2}\left\{\text{Tr}\left(\Sigma_2^{-1}(\mathbb{E}_{X\sim p}\left[(X-\mu_1)(X-\mu_1)^T+(X-\mu_1)(\mu_1-\mu_2)^T+(\mu_1-\mu_2)(X-\mu_1)^T+(\mu_1-\mu_2)(\mu_1-\mu_2)^T\right])\right)\right\} \\ =&\frac{1}{2}\left\{\text{Tr}\left(\Sigma_2^{-1}\left(\Sigma_1+\underbrace{\mathbb{E}_{X\sim p}[X-\mu_1]}_{=\mathbf{0}}(\mu_1-\mu_2)^T+(\mu_1-\mu_2)\underbrace{\mathbb{E}_{X\sim p}[(X-\mu_1)^T]}_{=\mathbf{0}}+(\mu_1-\mu_2)(\mu_1-\mu_2)^T\right)\right)\right\} \\ =&\frac{1}{2}\left\{\text{Tr}(\Sigma_2^{-1}\Sigma_1)+\text{Tr}\left(\Sigma_2^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T\right)\right\} \\ =&\frac{1}{2}\left(\text{Tr}(\Sigma_2^{-1}\Sigma_1)+\text{Tr}\left(\underbrace{(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)}_{\in\mathbb{R}}\right)\right) \\ =&\frac{1}{2}\,\text{Tr}(\Sigma_2^{-1}\Sigma_1)+\frac{1}{2}(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) \end{aligned}$$

Putting all terms together,

$$D_{KL}(p||q) = \frac{1}{2}\left(\log\frac{\text{det}(\Sigma_2)}{\text{det}(\Sigma_1)} - d + \text{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right)$$

[3] the calculation below relies on the cyclic property of trace.

B Marginal Distribution of Multivariate Gaussian

Let $X \sim \mathcal{N}(\mu_X, \Sigma_X) \in \mathbb{R}^{d_1}$ and $Y|X = x \sim \mathcal{N}(Ax + b, \Sigma_{Y|X}) \in \mathbb{R}^{d_2}$. We want to find the marginal of Y [4].

Theorem B.1. Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$, where $X_1 \in \mathbb{R}^{d_1}, X_2 \in \mathbb{R}^{d_2}$. Then given $X_2 = x_2$, the conditional distribution of X_1 is $(X_1|X_2) \sim \mathcal{N}(\mu_{1|2}, \Sigma_{11|2})$, where $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Proof. First, perform a nonsingular linear transformation:

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = BX = \begin{bmatrix} I_{d_1} & -\Sigma_{12}\Sigma_{22}^{-1} \\ \mathbf{O} & I_{d_2} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{bmatrix}$$

Hence, the mean and covariance matrix of Z are

$$\mu_Z = BX = \begin{bmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{bmatrix}$$

$$\Sigma_Z = B\Sigma B^T = \begin{bmatrix} I_{d_1} & -\Sigma_{12}\Sigma_{22}^{-1} \\ \mathbf{O} & I_{d_2} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I_{d_1} & \mathbf{O} \\ -\Sigma_{22}^{-1}\Sigma_{21} & I_{d_2} \end{bmatrix} = \begin{bmatrix} \underbrace{\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}_{:=\Sigma_{11|2}} & \mathbf{O} \\ \mathbf{O} & \Sigma_{22} \end{bmatrix}$$

Hence, $Z \sim \mathcal{N}(\mu_Z, \Sigma_Z)$. Also, Z_1 and Z_2 are independent because Σ_Z is diagonal. Then, we can write the joint densities of X and Z below [5]:

$$\begin{aligned} g(z_1, z_2) &= g_1(z_1)g_2(z_2) = g_1(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2)f_2(x_2), \\ f(x_1, x_2) &= g(z_1, z_2)\,|\det(\mathcal{J}_z)| \quad \mathcal{J}_z = B \text{ is the Jacobian matrix} \\ &= g_1(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2)f_2(x_2) \cdot 1 \end{aligned}$$

By definition of conditional distribution, the density of $X_1|X_2 = x_2$ is:

$$\begin{aligned} f_1(x_1|x_2) &= \frac{f(x_1, x_2)}{f_2(x_2)} = g_1(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2) \\ &= \frac{1}{(2\pi)^{\frac{d_1}{2}}\sqrt{\det(\Sigma_{11|2})}} \exp\{-\frac{1}{2}(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2 - (\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2))^T\Sigma_{11|2}^{-1} \\ &\qquad \underbrace{(x_1 - \Sigma_{12}\Sigma_{22}^{-1}x_2 - (\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2)\}}_{=x_1 - (\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)) := \mu_{1|2}} \\ &= \frac{1}{(2\pi)^{\frac{d_1}{2}}\sqrt{\det(\Sigma_{11|2})}} \exp\left\{-\frac{1}{2}(x_1 - \mu_{1|2})^T\Sigma_{11|2}^{-1}(x_1 - \mu_{1|2})\right\} \sim \mathcal{N}(\mu_{1|2}, \Sigma_{11|2}) \end{aligned}$$

□

Denote $Z = \begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} AX + b \\ X \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$. We know that $\mu_2 = \mu_X, \Sigma_{22} = \Sigma_X$, and also $\Sigma_{12} = \text{Cov}(X, AX + b) = A\,\text{Cov}(X, X) = A\Sigma_X, \Sigma_{21} = \Sigma_{12}^T = \Sigma_X A^T$.

[4] The following theorem is from this textbook [1], Theorem 2.3.2, pp.31-32

[5] The determinant of upper/lower-triangular matrix is product of the main diagonal, so $\det(B) = 1$.

Then we have, $\forall x \in \mathbb{R}^{d_1}$,

$$\begin{cases} Ax + b = \mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x - \mu_2) \\ \Sigma_{Y|X} = \Sigma_{22|1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{cases}$$

To find the marginal distribution of Y, we only need to solve for μ_1 and Σ_{11}:

$$\begin{aligned} Ax + b = \Sigma_{12}\Sigma_{22}^{-1}(x - \mu_2) = \mu_1 + A\Sigma_X\Sigma_X^{-1}(x - \mu_2) &\implies \mu_1 = A\mu_X + b \\ \Sigma_{Y|X} = \Sigma_{11} - A\Sigma_X\Sigma_X^{-1}\Sigma_X A^T &\implies \Sigma_{11} = \Sigma_{Y|X} + A\Sigma_X A^T \end{aligned}$$

Therefore, the marginal distribution of Y is $Y \sim \mathcal{N}(A\mu_X + b, \Sigma_{Y|X} + A\Sigma_X A^T)$.