Magda's 3rd set of technical notes

Last update: January 15, 2025

This is the 3rd set of my technical notes on various ML topics. I started writing the 1st set when beginning my PhD and the 2nd set when starting my PostDoc. Now I am a professor and have even less time than before but it feels that picking up at least some of the good old habits might help in bringing me back from admin to research. Both of the previous docs are available in my GitHub repo `https://github.com/mgswiss15/technotes`.

The general purpose of the notes is to help me understand better the selected topics by re-explaining (*re-* because these have been explained elsewhere many times), and to have a reference and possibly reusable material for later.

This is a working document not meant to be polished. There may be typos and other editing errors. Technical errors mean that I didn't quite understand something which I unfortunately cannot rule out.

Contents

1 Basics of diffusion models

Using: Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851
Lilian Weng. *What are Diffusion Models?* Section: posts. 2021. URL: `https://lilianweng.github.io/posts/2021-07-11-diffusion-models/` (visited on 12/17/2023).

1.1 Reverse process

We have true data coming from an unknown underlying distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$.

We assume a latent variable model $p_\theta(\mathbf{x}_0) \approx q(\mathbf{x}_0)$ approximating the true distribution $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_0|\mathbf{x}_{1:T})p_\theta(\mathbf{x}_{1:T})\,d\mathbf{x}_{1:T}$. For this we assume a learned *prior* $p_\theta(\mathbf{x}_{1:T})$ with Markov chain with Gaussian transitions (the means and variances are learned):

$$\begin{aligned} p_\theta(\mathbf{x}_{1:T}) &= p(\mathbf{x}_T)\prod_{t=2}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \\ p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &= \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \\ p(\mathbf{x}_T) &= \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \boldsymbol{I}) \quad . \end{aligned}$$

The complete joint distribution $p_\theta(\mathbf{x}_{0:T})$ is called the *reverse process*.

1.2 Variational bound

Follow the logic of importance sampling of VAE (see technical notes 2019): We could start maximizing the likelihood $p_\theta(\mathbf{x}_0)$ directly from

$$p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{1:T})p_\theta(\mathbf{x}_0|\mathbf{x}_{1:T})\,d\mathbf{x}_{1:T} \tag{1.1}$$

by sampling from the prior $p_\theta(\mathbf{x}_{1:T})$. Same as always, this would take very long cause the prior samples won't be very informative for the true data and won't give enough information for the training.

We could instead sample from the posterior $p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0) = \frac{p_\theta(\mathbf{x}_{0:T})}{p_\theta(\mathbf{x}_0)}$ using the *importance sampling* which should be more informative

$$p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{1:T})p_\theta(\mathbf{x}_0|\mathbf{x}_{1:T})}{p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)}\,d\mathbf{x}_{1:T} \tag{1.2}$$

The problem as always is that the posterior is intractable due to the unknown evidence $p_\theta(\mathbf{x}_0)$.

Hence we need an approximation instead $q(\mathbf{x}_{1:T}|\mathbf{x}_0) \approx p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)$ so that

$$p_\theta(\mathbf{x}_0) = \int q(\mathbf{x}_{1:T}|\mathbf{x}_0)\frac{p_\theta(\mathbf{x}_{1:T})p_\theta(\mathbf{x}_0|\mathbf{x}_{1:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\,d\mathbf{x}_{1:T} \tag{1.3}$$

This indeed is very similar to a VAE when we indicate the whole sequence $\mathbf{x}_{1:T}$ as a single latent variable $\mathbf{z}$.

We can now maximize the model likelihood $p_\theta(\mathbf{x}_0)$ by sampling the latent variable $\mathbf{x}_{1:T} \sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ from the approximate posteriors which can be seen as the *encoder*. We also have the learned *decoder* $p_\theta(\mathbf{x}_0|\mathbf{x}_{1:T})$ and a prior $p_\theta(\mathbf{x}_{1:T})$ which in this case is learned.

We could train the model from the classical variational bound on the log likelihood

$$\log p_\theta(\mathbf{x}_0) = \log \int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \, d\mathbf{x}_{1:T} \geq \int q(\mathbf{x}_{1:T}|\mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \, d\mathbf{x}_{1:T} \tag{1.4}$$

VAE-like minimization problem: Classically we minimize the negative log likelihood. The objective is thus the minimization of the variational bound

$$\begin{aligned}
\mathcal{L}(\mathbf{x}_0) &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} - \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T-1}|\mathbf{x}_T) p_\theta(\mathbf{x}_T)} \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \Big[- \log p_\theta(\mathbf{x}_T) - \log \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) + \log q(\mathbf{x}_{1:T}|\mathbf{x}_0) \Big]
\end{aligned} \tag{1.5}$$

1.3 Forward process

For the approximate posterior we assume again a Markov chain but now in the other direction as a *forward process* and with known Gaussian transitions with a variance schedule $\beta_1, \ldots, \beta_T$

$$\begin{aligned}
q(\mathbf{x}_{1:T}|\mathbf{x}_0) &= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \\
q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}) \ .
\end{aligned}$$

VAE-like minimization with forward process: The minimization problem (1.5) can now be written as

$$\begin{aligned}
\mathcal{L}(\mathbf{x}_0) &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \log \frac{\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[- \log p(\mathbf{x}_T) + \log \prod_{t=1}^{T} \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[- \log p(\mathbf{x}_T) + \sum_{t=1}^{T} \log \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \right]
\end{aligned} \tag{1.6}$$

Sampling from forward process: We can sample from forward process $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ recursively through $\mathbf{x}_t \sim \mathcal{N}(\sqrt{1-\beta_t})\mathbf{x}_{t-1}, \beta_t \boldsymbol{I})$ by fixing the previous value and sampling a noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon \quad \mathbb{E}(\mathbf{x}_t) = \sqrt{1-\beta_t}\mathbf{x}_{t-1} \quad \mathrm{Var}(\mathbf{x}_t) = \beta_t \boldsymbol{I} \tag{1.7}$$

Let us indicate $\alpha_t = 1 - \beta_t$ and hence $\beta_t = 1 - \alpha_t$

We then get

$$\begin{aligned}
\mathbf{x}_1 &\sim \mathcal{N}(\sqrt{\alpha_1}\mathbf{x}_0, (1-\alpha_1)\boldsymbol{I}) \\
\mathbf{x}_1 &= \sqrt{\alpha_1}\mathbf{x}_0 + \sqrt{1-\alpha_1}\epsilon_0
\end{aligned}$$

$$
\begin{aligned}
\mathbf{x}_2 &\sim \mathcal{N}(\sqrt{\alpha_2}\mathbf{x}_1, (1-\alpha_2)\boldsymbol{I}) \\
\mathbf{x}_2 &= \sqrt{\alpha_2}\mathbf{x}_1 + \sqrt{1-\alpha_2}\epsilon_1 \\
&= \sqrt{\alpha_2}(\sqrt{\alpha_1}\mathbf{x}_0 + \sqrt{1-\alpha_1}\epsilon_0) + \sqrt{1-\alpha_2}\epsilon_1 \\
&= \sqrt{\alpha_1\alpha_2}\mathbf{x}_0 + \sqrt{\alpha_2 - \alpha_1\alpha_2}\epsilon_0 + \sqrt{1-\alpha_2}\epsilon_1 \\
\mathbb{E}(\mathbf{x}_2) &= \sqrt{\alpha_1\alpha_2}\mathbf{x}_0 \\
\mathrm{Var}(\mathbf{x}_2) &= (\alpha_2 - \alpha_1\alpha_2 + 1 - \alpha_2)\boldsymbol{I} = (1-\alpha_1\alpha_2)\boldsymbol{I} \\
\mathbf{x}_2 &= \sqrt{\alpha_1\alpha_2}\mathbf{x}_0 + \sqrt{1-\alpha_1\alpha_2}\epsilon \\
\mathbf{x}_2 &\sim \mathcal{N}(\sqrt{\alpha_1\alpha_2}\mathbf{x}_0, (1-\alpha_1\alpha_2)\boldsymbol{I})
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{x}_3 &\sim \mathcal{N}(\sqrt{\alpha_3}\mathbf{x}_2, (1-\alpha_3)\boldsymbol{I}) \\
\mathbf{x}_2 &= \sqrt{\alpha_3}\mathbf{x}_2 + \sqrt{1-\alpha_3}\epsilon_2 \\
&= \sqrt{\alpha_3}(\sqrt{\alpha_1\alpha_2}\mathbf{x}_0 + \sqrt{\alpha_2 - \alpha_1\alpha_2}\epsilon_0 + \sqrt{1-\alpha_2}\epsilon_1) + \sqrt{1-\alpha_3}\epsilon_2 \\
&= \sqrt{\alpha_1\alpha_2\alpha_3}\mathbf{x}_0 + \sqrt{\alpha_2\alpha_3 - \alpha_1\alpha_2\alpha_3}\epsilon_0 + \sqrt{\alpha_3 - \alpha_2\alpha_3}\epsilon_1 + \sqrt{1-\alpha_3}\epsilon_2 \\
\mathbb{E}(\mathbf{x}_2) &= \sqrt{\alpha_1\alpha_2\alpha_3}\mathbf{x}_0 \\
\mathrm{Var}(\mathbf{x}_2) &= (\alpha_2\alpha_3 - \alpha_1\alpha_2\alpha_3 + \alpha_3 - \alpha_2\alpha_3 + 1 - \alpha_3)\boldsymbol{I} = (1-\alpha_1\alpha_2\alpha_3)\boldsymbol{I} \\
\mathbf{x}_2 &= \sqrt{\alpha_1\alpha_2\alpha_3}\mathbf{x}_0 + \sqrt{1-\alpha_1\alpha_2\alpha_3}\epsilon \\
\mathbf{x}_2 &\sim \mathcal{N}(\sqrt{\alpha_1\alpha_2\alpha_3}\mathbf{x}_0, (1-\alpha_1\alpha_2\alpha_3)\boldsymbol{I})
\end{aligned}
$$

and in general

$$
\begin{aligned}
\mathbf{x}_t &\sim \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\boldsymbol{I}) \\
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1} \\
\mathbf{x}_t &= \sqrt{\prod_{s=1}^{t}\alpha_s}\mathbf{x}_0 + \sqrt{1-\prod_{s=1}^{t}\alpha_s}\epsilon \\
\mathbf{x}_t &= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \\
\mathbf{x}_t &\sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}) = q(\mathbf{x}_t|\mathbf{x}_0)
\end{aligned} \tag{1.8}
$$

In summary, instead of fixing the variance schedule $\beta_1 \dots \beta_T$ and sampling from the forward process recursively

$$
\mathbf{x}_t \sim \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\boldsymbol{I}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}) \quad \text{via} \quad \mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{1.9}
$$

we can fix the schedule $\bar{\alpha}_1 \dots \bar{\alpha}_T$ and sample arbitrary timestep directly from

$$
\mathbf{x}_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}) = q(\mathbf{x}_t|\mathbf{x}_0) \quad \text{via} \quad \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{1.10}
$$

Minimization with x_0 conditioning: We can further play with the variational bound (1.6)

$$
\begin{aligned}
\mathcal{L}(\mathbf{x}_0) &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[-\log p(\mathbf{x}_T) + \sum_{t=1}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[-\log p(\mathbf{x}_T) + \sum_{t=2}^{T}\log\frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \log\frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}\right]
\end{aligned} \tag{1.11}
$$

Now we use the following for the forward process

$$\begin{aligned} q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0) &= q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) q(\mathbf{x}_{t-1}|\mathbf{x}_0) q(\mathbf{x}_0) \\ &= q(\mathbf{x}_t|\mathbf{x}_{t-1}) q(\mathbf{x}_{t-1}|\mathbf{x}_0) q(\mathbf{x}_0) \quad \text{(Markov assumption on forward process)} \\ q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0) &= q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t|\mathbf{x}_0) q(\mathbf{x}_0) \end{aligned} \tag{1.12}$$

so that

$$\begin{aligned} q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) \\ &= \frac{q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0) q(\mathbf{x}_0)} \\ &= \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t|\mathbf{x}_0) q(\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0) q(\mathbf{x}_0)} \end{aligned} \tag{1.13}$$

to rewrite the minimization as

$$\mathcal{L}(\mathbf{x}_0) = \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[-\log p(\mathbf{x}_T) + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)} \right] \tag{1.14}$$

We then observe that

$$\begin{aligned} \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)} &= \log \prod_{t=2}^{T} \frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)} \\ &= \log \frac{q(\mathbf{x}_2|\mathbf{x}_0)}{q(\mathbf{x}_1|\mathbf{x}_0)} \frac{q(\mathbf{x}_3|\mathbf{x}_0)}{q(\mathbf{x}_2|\mathbf{x}_0)} \frac{q(\mathbf{x}_4|\mathbf{x}_0)}{q(\mathbf{x}_3|\mathbf{x}_0)} \cdots \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{q(\mathbf{x}_T - 1|\mathbf{x}_0)} \\ &= \log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{q(\mathbf{x}_1|\mathbf{x}_0)} \end{aligned} \tag{1.15}$$

and hence get for a single sample

$$\begin{aligned} \mathcal{L}(\mathbf{x}_0) &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[-\log p(\mathbf{x}_T) + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{q(\mathbf{x}_1|\mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)} \right] \\ &= \mathbb{E}_{q(\mathbf{x}_T|\mathbf{x}_0)} \log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p(\mathbf{x}_T)} - \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)} \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) + \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \end{aligned} \tag{1.16}$$

and in expectation for the complete data set

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q(\mathbf{x}_0)} \mathcal{L}(\mathbf{x}_0) \\ &= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[-\log p(\mathbf{x}_T) + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} + \log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{q(\mathbf{x}_1|\mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)} \right] \\ &= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[-\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] \\ &= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\underbrace{D_{\mathrm{KL}}\left(q(\mathbf{x}_T|\mathbf{x}_0) \,||\, p(\mathbf{x}_T)\right)}_{\mathcal{L}_T} + \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,||\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\right)}_{\mathcal{L}_{t-1}} \underbrace{- \log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{\mathcal{L}_0} \right] \end{aligned} \tag{1.17}$$

What is the missing term to complete the bound (see VAE)?

The forward process posterior $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ conditioned on $\mathbf{x}_0$ is tractable and can be compared to the learned reversed process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \tag{1.18}$$

Pdf of exponential family distributions can be represented in a form

$$p(\mathbf{x};\boldsymbol{\eta}) = h(\mathbf{x}) \exp\left(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{T}(\mathbf{x}) - \boldsymbol{A}(\boldsymbol{\eta})\right) \quad . \tag{1.19}$$

For multivariate Guassian distribution we get Escudero 2020

$$\begin{aligned}
\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= (2\pi)^{-k/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \\
&= \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma})\right) \\
&= \exp\left(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}) + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\mathbf{x}^{\mathrm{T}}) - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma})\right) \\
&= \exp\left(-\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}\mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{\mathrm{T}}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma})\right)
\end{aligned}$$

From which we have

$$\begin{aligned}
\boldsymbol{\eta} &= \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix} \\
T(\mathbf{x}) &= \begin{bmatrix} \mathbf{x} \\ \mathrm{vec}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) \end{bmatrix} \\
\boldsymbol{A}(\boldsymbol{\eta}) &= -\frac{1}{2}\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + k\log(2\pi) + \log\det(\boldsymbol{\Sigma})\right)
\end{aligned}$$

For the forward process we have

$$\begin{aligned}
q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\boldsymbol{I}) \\
&= \exp\left(-\frac{1}{2\beta_t}\mathbf{x}_t^{\mathrm{T}}\mathbf{x}_t + \frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t^{\mathrm{T}}\mathbf{x}_{t-1} - \frac{\alpha_t}{2\beta_t}\mathbf{x}_{t-1}^{\mathrm{T}}\mathbf{x}_{t-1} - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log k\beta_t\right)
\end{aligned}$$

$$\begin{aligned}
q(\mathbf{x}_t|\mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}) \\
&= \exp\left(-\frac{1}{2(1-\bar{\alpha}_t)}\mathbf{x}_t^{\mathrm{T}}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_t}}{(1-\bar{\alpha}_t)}\mathbf{x}_t^{\mathrm{T}}\mathbf{x}_0 - \frac{\bar{\alpha}_t}{2(1-\bar{\alpha}_t)}\mathbf{x}_0^{\mathrm{T}}\mathbf{x}_0 - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log k(1-\bar{\alpha}_t)\right)
\end{aligned}$$

$$\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0, (1-\bar{\alpha}_{t-1})\boldsymbol{I}) \\
&= \exp\left(-\frac{1}{2(1-\bar{\alpha}_{t-1})}\mathbf{x}_{t-1}^{\mathrm{T}}\mathbf{x}_{t-1} + \frac{\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_{t-1})}\mathbf{x}_{t-1}^{\mathrm{T}}\mathbf{x}_0 - \frac{\bar{\alpha}_{t-1}}{2(1-\bar{\alpha}_{t-1})}\mathbf{x}_0^{\mathrm{T}}\mathbf{x}_0 - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log k(1-\bar{\alpha}_{t-1})\right)
\end{aligned}$$

And hence

$$\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &= \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\
&= \exp\left(\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t^{\mathrm{T}}\mathbf{x}_{t-1} - \frac{\alpha_t}{2\beta_t}\mathbf{x}_{t-1}^{\mathrm{T}}\mathbf{x}_{t-1} - \frac{1}{2(1-\bar{\alpha}_{t-1})}\mathbf{x}_{t-1}^{\mathrm{T}}\mathbf{x}_{t-1} + \frac{\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_{t-1})}\mathbf{x}_{t-1}^{\mathrm{T}}\mathbf{x}_0 - \boldsymbol{A}(\boldsymbol{\eta})\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\alpha_t}{\beta_t} + \frac{1}{(1-\bar{\alpha}_{t-1})}\right)\mathbf{x}_{t-1}^{\mathrm{T}}\mathbf{x}_{t-1} + \mathbf{x}_{t-1}^{\mathrm{T}}\left(\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_{t-1})}\mathbf{x}_0\right) - \boldsymbol{A}(\boldsymbol{\eta})\right) \quad .
\end{aligned}$$

where $\boldsymbol{A}(\boldsymbol{\eta})$ is the log-partition function (cummulant) contains all the normalizing terms not depending on $\mathbf{x}_{t-1}$.

From this we have that the covariance of the distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ is

$$\boldsymbol{\Sigma} = \left(\frac{\alpha_t}{\beta_t} + \frac{1}{(1-\bar{\alpha}_{t-1})}\right)^{-1} \boldsymbol{I} = \left(\frac{\beta_t + \alpha_t - \bar{\alpha}_t}{\beta_t(1-\bar{\alpha}_{t-1})}\right)^{-1} \boldsymbol{I} = \frac{\beta_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{I} = \bar{\beta}_t \boldsymbol{I} \ ,$$

and the mean is

$$\begin{aligned}\boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0) &= \left(\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_{t-1})}\mathbf{x}_0\right)\frac{\beta_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \\ &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_t)}\mathbf{x}_0\end{aligned}$$

So that $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0), \bar{\beta}_t \boldsymbol{I})$.

From (1.8) we know that we can sample $\mathbf{x}_t$ in the forward diffusion directly from $\mathbf{x}_0$ as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$ with $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Hence we can also recover the $\mathbf{x}_0$ from the sample (if we know the noise) as $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t\right)$. This also means that the posterior is

$$q(\mathbf{x}_0|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_0; \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{x}_t, \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\boldsymbol{I}) \tag{1.20}$$

For the mean of the distribution we thus get in terms of the known error ϵ_t

$$\begin{aligned}\boldsymbol{\mu}(\mathbf{x}_t, \epsilon_t) &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_t)}\mathbf{x}_0 \\ &= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{(1-\bar{\alpha}_t)}\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t\right) \\ &= \frac{\alpha_t(1-\bar{\alpha}_{t-1}) + (1-\alpha_t)}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\mathbf{x}_t - \frac{(1-\alpha_t)}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_t \\ &= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{(1-\alpha_t)}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right)\end{aligned} \tag{1.21}$$

So that the posterior of the forward process can be conditionend on the known error sample $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \epsilon_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}(\mathbf{x}_t, \epsilon_t), \bar{\beta}_t \boldsymbol{I})$.

2 Simplifying the loss function - denoising autoencoder

The $\mathcal{L}_T$ term in (1.17) has no learnable parameters. The prior $p(\mathbf{x}_T)$ is standard normal and $q(\mathbf{x}_T|\mathbf{x}_0)$ parameters depend only on the forward variance schedule through (1.8). It can just be dropped from the objective.

General the KL divergence for two multivariate Gaussians is as follows

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \,||\, \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)) = \frac{1}{2}\left(\log\frac{\det(\boldsymbol{\Sigma}_p)}{\det(\boldsymbol{\Sigma}_q)} - k + \mathrm{Tr}\left(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q\right) + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)\right)$$

For the $\mathcal{L}_{t-1}$ terms the posterior of the forward process $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0), \bar{\beta}_t \boldsymbol{I})$ has no learnable parameters. We fix the variance in the reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$ so that $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \boldsymbol{I}$, where σ_t^2 is a known (not learnable) function of the forward variance

schedule. The minimization of the KL divergence hence simplifies to:

$$\arg\min_{\theta} \mathbb{E}_{q(\mathbf{x}_0,\mathbf{x}_t)} D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \;||\; p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)\right)$$

is equivalent to

$$\arg\min_{\theta} \mathbb{E}_{q(\mathbf{x}_0,\mathbf{x}_t)} \frac{1}{2\sigma_t^2} \left\|\boldsymbol{\mu}(\mathbf{x}_t,\mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\right\|_2^2 \tag{2.1}$$

and hence we can train $\boldsymbol{\mu}_\theta$ to approximate the mean of the forward process posterior.

However, we know from (1.21) that mean of the forward posterior can be written with respect to $\mathbf{x}_t$ and the noise which ϵ_t which was used to generate if from $\mathbf{x}_0$.

$$\boldsymbol{\mu}(\mathbf{x}_t,\epsilon_t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{(1-\alpha_t)}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right)$$

We can therefore choose the parameterization for the mean of the reverse process as

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{(1-\alpha_t)}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\right)$$

In this parametrization is the reverse process

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{(1-\alpha_t)}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\right), \sigma_t^2\boldsymbol{I})$$

and we can sample $\mathbf{x}_{t-1}$ as

$$\mathbf{x}_{t-1} = \left(\mathbf{x}_t - \frac{(1-\alpha_t)}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\right) + \sigma_t\mathbf{z}, \qquad \mathbf{z}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})$$

Plugging these back to the optimisation from (2) we get

$$\begin{aligned}
\left\|\boldsymbol{\mu}(\mathbf{x}_t,\mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\right\|_2^2 &= \left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{(1-\alpha_t)}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right) - \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{(1-\alpha_t)}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\right)\right\|_2^2 \\
&= \left\|\frac{(1-\alpha_t)}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\left(\epsilon_t - \epsilon_\theta(\mathbf{x}_t,t)\right)\right\|_2^2 \\
&= \frac{(1-\alpha_t)^2}{\alpha_t(1-\bar{\alpha}_t)}\left\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\right\|_2^2
\end{aligned}$$

This means that the $\mathcal{L}_{t-1}$ terms of the objective boil down to

$$\mathcal{L}_{t-1}: \arg\min_{\theta} \mathbb{E}_{q(\mathbf{x}_0)\mathcal{N}(\epsilon_t;\mathbf{0},\boldsymbol{I})} \frac{(1-\alpha_t)^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\epsilon_t - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t}_{\mathbf{x}_t}, t)\right\|_2^2,$$

whereby the diffusion model is trained to predict the noise from the noised image and the corresponding timestamp.

In Ho, Jain, and Abbeel 2020 they model $p_\theta(\mathbf{x}_0|\mathbf{x}_1)$ as independent discrete decoder over the image pixels. Check the paper for details.

They also found that the don't need to re-weight the loss terms so that in the end we have a simple objective

$$\mathcal{L}_{simple}: \arg\min_{\theta} \sum_{t=1}^{T}\left\|\epsilon_t - \epsilon_\theta(\underbrace{\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t}_{\mathbf{x}_t}, t)\right\|_2^2,$$

where the $p_\theta(\mathbf{x}_0|\mathbf{x}_1)$ has been subsumed into the loss.

I don't quite see, how this happens but seems not very critical.

The final point here is that we can use the trained model ϵ_θ to predict the original image $\widehat{\mathbf{x}}_0$ from the noised image $\mathbf{x}_t$ and the timestamp as

$$\mathbf{x}_0 \approx \widehat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)\right) = \boldsymbol{u}_\theta(\mathbf{x}_t, t) \tag{2.2}$$

and we call this function $\boldsymbol{u}_\theta(\mathbf{x}_t, t)$.

Some comments on this It is important to understand the properties of this approximator. When $\mathbf{x}_t$ is the result of the forward process, the mean of this is

$$\begin{aligned}\mathbb{E}(\widehat{\mathbf{x}}_0) &= \mathbb{E}\left[\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\right)\right] \\ &= \mathbf{x}_0 - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\mathbb{E}(\epsilon_\theta)\end{aligned}$$

Hence the bias is a function of the bias of the predictor ϵ_θ.

The variance

$$\begin{aligned}\mathrm{Var}(\widehat{\mathbf{x}}_0) &= \mathrm{Var}\left[\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\right)\right] \\ &= \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}(\boldsymbol{I} + \mathrm{Var}(\epsilon_\theta)) \ .\end{aligned}$$

It can be expected that the variance $\mathrm{Var}(\epsilon_\theta)$ is larger for bigger t (further away from the original image). The variance schedule has an effect on this through $\bar{\alpha}_t$.

I can't think it through but probably worse exploring a bit more.

When $\mathbf{x}_t$ comes from the reverse process the moments are not the same. Again, I cannot think it through now.

3 Basics of classifier guidance

Using: Prafulla Dhariwal and Alexander Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794

3.1 Classifier trained on noisy images

Let's assume a classifier $p_\phi(y|\mathbf{x}_t)$ trained on the noisy images $\mathbf{x}_t$ and use the gradients $\nabla_{\mathbf{x}_t} p_\phi(y|\mathbf{x}_t)$ to guide the diffusion sampling.

4 Denoising diffusion implicit models

Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: International Conference on Learning Representations. 2021

5 Background DDPM

We start the same as in the DDPM of Ho, Jain, and Abbeel 2020.

Goal is to learn model $p_\theta(\mathbf{x}_0) \approx q(\mathbf{x}_0)$ approximating the true data distribution. We formulate the model as latent variable with latents $\mathbf{x}_{1:T}$

$$p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T} = \int p_\theta(\mathbf{x}_{1:T})\, p_\theta(\mathbf{x}_0 \mid \mathbf{x}_{1:T})\, d\mathbf{x}_{1:T} = \int p_\theta(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\, d\mathbf{x}_{1:T} \ . \tag{5.1}$$

This is the diffusion *reverse or generative process.*

The posterior of the latents is

$$p_\theta(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \frac{p_\theta(\mathbf{x}_{1:T})\, p_\theta(\mathbf{x}_0 \mid \mathbf{x}_{1:T})}{p_\theta(\mathbf{x}_0)} = \frac{p_\theta(\mathbf{x}_{0:T})}{p_\theta(\mathbf{x}_0)} \ . \tag{5.2}$$

We approximate the posterior by a fixed *encoder, forward process or inference distribution* $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \approx p_\theta(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$.

We learn the model parameters θ by maximizing the ELBO

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{x}_0)} \left[\log p_\theta(\mathbf{x}_0)\right] &= \mathbb{E}_{q(\mathbf{x}_0)} \left[\log \int p_\theta(\mathbf{x}_{0:T})\, d\mathbf{x}_{1:T}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_0)} \left[\log \int q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} d\mathbf{x}_{1:T}\right] \\
&\geq \mathbb{E}_{q(\mathbf{x}_0)} \left[\int q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} d\mathbf{x}_{1:T}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_0)} \mathbb{E}_{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} \\
&= \mathbb{E}_{q(\mathbf{x}_{0:T})} \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} = \text{ELBO} \ ,
\end{aligned} \tag{5.3}$$

which is obviously equivalent to minimizing the negative ELBO

$$\arg\min_\theta -\text{ELBO} = \arg\min_\theta \mathbb{E}_{q(\mathbf{x}_{0:T})} - \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} \ . \tag{5.4}$$

In DDPM the forward process was fixed as a Markov chain, such that

$$q(\mathbf{x}_{0:T}) = q(\mathbf{x}_0) \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1}), \quad q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \sim \mathcal{N}\left(\mathbf{x}_t; \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\mathbf{x}_{t-1}, \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right) \boldsymbol{I}\right) \ , \tag{5.5}$$

with the following link to the initial notation of Ho, Jain, and Abbeel 2020

$$\alpha_t = \prod_{s=1}^{t} (1 - \beta_t), \quad \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right) = \beta_t, \quad \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} = \sqrt{1 - \beta_t} \ . \tag{5.6}$$

By the same logic as in Ho, Jain, and Abbeel 2020 it also holds that

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \int q(\mathbf{x}_{1:t} \mid \mathbf{x}_0) d\mathbf{x}_{1:(t-1)} = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\boldsymbol{I}\right) \ , \tag{5.7}$$

with $\lim_{t\to\infty} \alpha_t = 0$ and hence $\lim_{t\to\infty} q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
Observe that by the Markov assumption on the forward process we have

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0) q(\mathbf{x}_0)} = \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t \mid \mathbf{x}_0) q(\mathbf{x}_0)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0) q(\mathbf{x}_0)} \ . \tag{5.8}$$

We can use it in the ELBO

$$\begin{aligned}
\text{ELBO} &= \mathbb{E}_{q(\mathbf{x}_{0:T})} \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} \\
&= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log p_\theta(\mathbf{x}_T) + \log \prod_{t=1}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_t \mid \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log p_\theta(\mathbf{x}_T) + \log \prod_{t=1}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t \mid \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log p_\theta(\mathbf{x}_T) + \log \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} + \log \frac{p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1)}{q(\mathbf{x}_T \mid \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1) + \log \prod_{t=2}^{T} \frac{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{0:T})} \left[\log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1) - \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{0:1})} \log p_\theta(\mathbf{x}_0 \mid \mathbf{x}_1) - \mathbb{E}_{q(\mathbf{x}_0)} \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,||\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\right) \ ,
\end{aligned} \tag{5.9}$$

where we assume that $p_\theta(\mathbf{x}_T) = q(\mathbf{x}_T \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and therefore drop it (also we cannot influence these by training so can be).
We further have from Ho, Jain, and Abbeel 2020

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1};\, \boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0), \bar{\beta}_t \boldsymbol{I}) \ , \tag{5.10}$$

where

$$\boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0) = \frac{\beta_t \sqrt{\alpha_{t-1}}}{(1-\alpha_t)} \mathbf{x}_0 + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t} \mathbf{x}_t \tag{5.11}$$

and

$$\bar{\beta}_t = \frac{\beta_t (1-\alpha_{t-1})}{1-\alpha_t} \tag{5.12}$$

When we put $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1};\, \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2 \boldsymbol{I}\right)$ we get for the KL divergences

$$D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,||\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\right) = \frac{1}{2\sigma_t^2} \|\boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|_2^2 \tag{5.13}$$

Using the fact that $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\boldsymbol{I}\right)$, we can sample $\mathbf{x}_t$ as

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{(1-\alpha_t)}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{5.14}$$

and hence after we have sampled ϵ we can recover $\mathbf{x}_0$ from $\mathbf{x}_t$ as

$$\mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{(1-\alpha_t)}\epsilon}{\sqrt{\alpha_t}} \ . \tag{5.15}$$

With this we can

$$\begin{aligned}
\boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\beta_t\sqrt{\alpha_{t-1}}}{(1-\alpha_t)}\mathbf{x}_0 + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{x}_t \\
&= \frac{\beta_t\sqrt{\alpha_{t-1}}}{(1-\alpha_t)\sqrt{\alpha_t}}\left(\mathbf{x}_t - \sqrt{(1-\alpha_t)}\epsilon\right) + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{x}_t \\
&= \frac{\beta_t}{(1-\alpha_t)\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \sqrt{(1-\alpha_t)}\epsilon\right) + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{x}_t \\
&= \frac{\beta_t + 1 - \beta_t - \alpha_t}{(1-\alpha_t)\sqrt{(1-\beta_t)}}\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\beta_t}\sqrt{(1-\alpha_t)}}\epsilon \\
&= \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{(1-\alpha_t)}}\epsilon\right) = \boldsymbol{\mu}(\mathbf{x}_t, \epsilon) \ .
\end{aligned} \tag{5.16}$$

We can set

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{(1-\alpha_t)}}\epsilon_\theta(\mathbf{x}_t, t)\right) \tag{5.17}$$

and therefore get

$$\begin{aligned}
\frac{1}{2\sigma_t^2}\|\boldsymbol{\mu}(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|_2^2 &= \frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{(1-\alpha_t)}}\epsilon\right) - \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{(1-\alpha_t)}}\epsilon_\theta(\mathbf{x}_t, t)\right)\right\|_2^2 \\
&= \frac{\beta_t^2}{2\sigma_t^2\,(1-\beta_t)(1-\alpha_t)}\left\|\epsilon - \epsilon_\theta(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{(1-\alpha_t)}\epsilon, t)\right\|_2^2 \ .
\end{aligned} \tag{5.18}$$

More generally, we can write the loss as

$$\sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{x}_0)}D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,||\, p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\right) = \sum_{t=2}^{T}\mathbb{E}_{q(\mathbf{x}_0)}\gamma_t\left\|\epsilon - \epsilon_\theta(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{(1-\alpha_t)}\epsilon, t)\right\|_2^2$$

$$Ł_\gamma(\epsilon_\theta) = \sum_{t=2}^{T}\gamma_t\mathbb{E}_{q(\mathbf{x}_0),\epsilon_t}\left\|\epsilon_t - \epsilon_\theta^t(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{(1-\alpha_t)}\epsilon_t)\right\|_2^2 \tag{5.19}$$

In the paper they start the sum from $t = 1$. I still do not quite understand the first step.

6 Moving onto DDIM

We take a different assumption for $q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)$ formulated as *reverse process*

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = q_\sigma(\mathbf{x}_T|\mathbf{x}_0)\prod_{t=2}^{T}q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \ , \tag{6.1}$$

where

$$q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\right) \tag{6.2}$$

and

$$q_\sigma(\mathbf{x}_T|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_T; \sqrt{\alpha_T}\mathbf{x}_0, (1-\alpha_T)\boldsymbol{I}\ .\right) \tag{6.3}$$

This is chosen so that it still holds (as in the DDPM) that

$$q_\sigma(\mathbf{x}_t \mid \mathbf{x}_0) = \int q(\mathbf{x}_{1:t} \mid \mathbf{x}_0) d\mathbf{x}_{1:(t-1)} = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\boldsymbol{I}\right)\ . \tag{6.4}$$

This hold by assumption for $q_\sigma(\mathbf{x}_T \mid \mathbf{x}_0)$. We can start from $t = T$ and then prove by induction that it holds for all t. We use marginalization formula

$$q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_0) = \int q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)\, q_\sigma(\mathbf{x}_t \mid \mathbf{x}_0) d\mathbf{x}_t\ . \tag{6.5}$$

The q_σ on the right side are both gaussians and Bisshop 2.115 says that

$$\begin{aligned} q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_0) &= \mathcal{N}\left(\mathbf{x}_{t-1};\ \sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\frac{\sqrt{\alpha_t}\mathbf{x}_0 - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I} + \frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}(1-\alpha_t)\boldsymbol{I}\right) \\ &= \mathcal{N}\left(\mathbf{x}_{t-1};\ \sqrt{\alpha_{t-1}}\mathbf{x}_0, (1-\alpha_{t-1})\boldsymbol{I}\right)\ . \end{aligned} \tag{6.6}$$

Though $q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$ depends on σ, $q_\sigma(\mathbf{x}_t \mid \mathbf{x}_0)$ actually does not.

The corresponding *forward process* is again Gaussian though I do not need the Markov assumption as in DDPM

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)q(\mathbf{x}_0)} = \frac{q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)q(\mathbf{x}_t \mid \mathbf{x}_0)q(\mathbf{x}_0)}{q(\mathbf{x}_{t-1} \mid \mathbf{x}_0)q(\mathbf{x}_0)}\ . \tag{6.7}$$

When $\sigma \to 0$ the reverse process is deterministic

$$\lim_{\sigma\to 0} q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}}\,\frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}},\ 0\boldsymbol{I}\right)\ , \tag{6.8}$$

so that

$$\mathbf{x}_{t-1} = \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\mathbf{x}_t - \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{\alpha_{t-1}}\mathbf{x}_0\ . \tag{6.9}$$

In consequence the forward process is also deterministic

$$q_\sigma(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_0) = \sqrt{\alpha_t}\mathbf{x}_0 - \sqrt{\frac{1-\alpha_t}{1-\alpha_{t-1}}}\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{\frac{1-\alpha_t}{1-\alpha_{t-1}}}\mathbf{x}_{t-1}\ . \tag{6.10}$$

As in DDPM we can sample $\mathbf{x}_t$ from the same distribution

$$\mathbf{x}_t = \sqrt{\alpha_t}\,\mathbf{x}_0 + \sqrt{1-\alpha_t}\,\epsilon_t \tag{6.11}$$

and similarly as before we can reverse this and predict the de-noised observation

$$\widehat{\mathbf{x}}_0(\mathbf{x}_t) = \frac{\mathbf{x}_t - \sqrt{(1-\alpha_t)}\,\epsilon_\theta^{(t)}(\mathbf{x}_t)}{\sqrt{\alpha_t}}\ . \tag{6.12}$$

We also have

$$\epsilon_\theta^{(t)}(\mathbf{x}_t) = \frac{\mathbf{x}_t - \sqrt{\alpha_t}\,\widehat{\mathbf{x}}_0(\mathbf{x}_t)}{\sqrt{(1-\alpha_t)}}\ . \tag{6.13}$$

Using this we can define the generative process starting from $p_\theta(\mathbf{x}_T) = N(0, I)$ and then

$$p_\theta^{(t)}(\mathbf{x}_{t-1}|\mathbf{x}_t) = q_\sigma(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \widehat{\mathbf{x}}_0(\mathbf{x}_t)) = \mathcal{N}\left(\mathbf{x}_{t-1}; \sqrt{\alpha_{t-1}}\,\widehat{\mathbf{x}}_0(\mathbf{x}_t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\frac{\mathbf{x}_t - \sqrt{\alpha_t}\,\widehat{\mathbf{x}}_0(\mathbf{x}_t)}{\sqrt{1-\alpha_t}},\ \sigma_t^2\boldsymbol{I}\right) \tag{6.14}$$

and

$$p_\theta^{(1)}(\mathbf{x}_0|\mathbf{x}_1) = \mathcal{N}\left(\mathbf{x}_0;\, \widehat{\mathbf{x}}_0(\mathbf{x}_1),\, \sigma_1^2 \boldsymbol{I}\right) \quad . \tag{6.15}$$

The ELBO is again

$$\begin{aligned}
\text{ELBO} &= \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[\log p_\theta(\mathbf{x}_T) + \log \prod_{t=2}^{T} \frac{p_\theta^{(t)}(\mathbf{x}_{t-1} \mid \mathbf{x}_t)}{q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)} + \log \frac{p_\theta^{(1)}(\mathbf{x}_0 \mid \mathbf{x}_1)}{q_\sigma(\mathbf{x}_T \mid \mathbf{x}_0)}\right] \\
&= \mathbb{E}_{q(\mathbf{x}_{0:1})} \log p_\theta^{(1)}(\mathbf{x}_0 \mid \mathbf{x}_1) - \mathbb{E}_{q(\mathbf{x}_0)} \sum_{t=2}^{T} D_{\mathrm{KL}}\left(q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,||\, p_\theta^{(t)}(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\right) \quad .
\end{aligned} \tag{6.16}$$

We look at

$$\begin{aligned}
&D_{\mathrm{KL}}\left(q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,||\, p_\theta^{(t)}(\mathbf{x}_{t-1} \mid \mathbf{x}_t)\right) \\
=&D_{\mathrm{KL}}\left(q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \,||\, q_\sigma(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \widehat{\mathbf{x}}_0(\mathbf{x}_t))\right) \\
=&\frac{1}{\sigma_t^2}\left\| \sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}} - \sqrt{\alpha_{t-1}}\widehat{\mathbf{x}}_0(\mathbf{x}_t)) - \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\frac{\mathbf{x}_t - \sqrt{\alpha_t}\widehat{\mathbf{x}}_0(\mathbf{x}_t))}{\sqrt{1-\alpha_t}} \right\|_2^2 \\
=&\frac{1}{\sigma_t^2}\left\| \sqrt{\alpha_{t-1}}(\mathbf{x}_0 - \widehat{\mathbf{x}}_0(\mathbf{x}_t)) - \frac{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\sqrt{\alpha_t}}{\sqrt{1-\alpha_t}}(\mathbf{x}_0 - \widehat{\mathbf{x}}_0(\mathbf{x}_t)) \right\|_2^2 \\
=&\frac{1}{\sigma_t^2}\left\| \frac{\sqrt{1-\alpha_t}\sqrt{\alpha_{t-1}} - \sqrt{1-\alpha_{t-1}-\sigma_t^2}\sqrt{\alpha_t}}{\sqrt{1-\alpha_t}}(\mathbf{x}_0 - \widehat{\mathbf{x}}_0(\mathbf{x}_t)) \right\|_2^2 \\
=&\frac{(\sqrt{1-\alpha_t}\sqrt{\alpha_{t-1}} - \sqrt{1-\alpha_{t-1}-\sigma_t^2}\sqrt{\alpha_t})^2}{\sigma_t^2(1-\alpha_t)} \left\|\mathbf{x}_0 - \widehat{\mathbf{x}}_0(\mathbf{x}_t)\right\|_2^2 \\
=&\frac{(\sqrt{1-\alpha_t}\sqrt{\alpha_{t-1}} - \sqrt{1-\alpha_{t-1}-\sigma_t^2}\sqrt{\alpha_t})^2}{\sigma_t^2(1-\alpha_t)} \left\| \frac{\mathbf{x}_t - \sqrt{(1-\alpha_t)}\,\epsilon_t}{\sqrt{\alpha_t}} - \frac{\mathbf{x}_t - \sqrt{(1-\alpha_t)}\,\epsilon_\theta^{(t)}(\mathbf{x}_t)}{\sqrt{\alpha_t}} \right\|_2^2 \\
=&\frac{(\sqrt{1-\alpha_t}\sqrt{\alpha_{t-1}} - \sqrt{1-\alpha_{t-1}-\sigma_t^2}\sqrt{\alpha_t})^2}{\sigma_t^2\alpha_t} \left\|\epsilon_t - \epsilon_\theta^{(t)}(\mathbf{x}_t)\right\|_2^2
\end{aligned} \tag{6.17}$$

This is up to the terms before the norm which do not depend on θ the same as in DDPM. Since we wish to maximize ELBO which is a sum of these KLs across t, the optimal solution is reached when each of the norms is minimized irrespective of the weighting.

7 Sampling from DDIM

We had before that

$$p_\theta^{(t)}(\mathbf{x}_{t-1}|\mathbf{x}_t) = q_\sigma(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \widehat{\mathbf{x}}_0(\mathbf{x}_t)) = \mathcal{N}\left(\mathbf{x}_{t-1}; \sqrt{\alpha_{t-1}}\,\widehat{\mathbf{x}}_0(\mathbf{x}_t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\frac{\mathbf{x}_t - \sqrt{\alpha_t}\,\widehat{\mathbf{x}}_0(\mathbf{x}_t)}{\sqrt{1-\alpha_t}},\, \sigma_t^2 \boldsymbol{I}\right) \tag{7.1}$$

and hence we can sample

$$\begin{aligned}
\mathbf{x}_{t-1} &= \sqrt{\alpha_{t-1}}\,\widehat{\mathbf{x}}_0(\mathbf{x}_t) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\frac{\mathbf{x}_t - \sqrt{\alpha_t}\,\widehat{\mathbf{x}}_0(\mathbf{x}_t)}{\sqrt{1-\alpha_t}} + \sigma_t\epsilon_t \\
&= \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{(1-\alpha_t)}\,\epsilon_\theta^{(t)}(\mathbf{x}_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\epsilon_\theta^{(t)}(\mathbf{x}_t) + \sigma_t\epsilon_t
\end{aligned} \tag{7.2}$$

References

[1] Prafulla Dhariwal and Alexander Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794 (cit. on p. 10).

[2] Mauro Camara Escudero. *Multivariate Normal as an Exponential Family Distribution*. Mauro Camara Escudero. 2020. URL: `https://maurocamaraescudero.netlify.app/post/multivariate-normal-as-an-exponential-family-distribution/` (visited on 12/17/2023) (cit. on p. 6).

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851 (cit. on pp. 2, 8, 11, 12).

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: International Conference on Learning Representations. 2021 (cit. on p. 11).

[5] Lilian Weng. *What are Diffusion Models?* Section: posts. 2021. URL: `https://lilianweng.github.io/posts/2021-07-11-diffusion-models/` (visited on 12/17/2023) (cit. on p. 2).

Index