From VAE to Diffusion Model

Varational Autoencoder (VAE)

In the normal auto-encoder (AE) model, for a data distribution $p(\boldsymbol{x})$, we first encode $\boldsymbol{x}$ using $q(\boldsymbol{z}|\boldsymbol{x})$ and decode it using $p(\boldsymbol{x}|\boldsymbol{z})$. We need to optimize the loglikelihood of $\boldsymbol{x}$ for a given encoding function $q(\boldsymbol{z}|\boldsymbol{x})$. This gives the following loss function:

$$\mathcal{L}_{\mathbf{a}e}(x) = \mathbf{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[p(\boldsymbol{x}|\boldsymbol{z})] + P_\alpha$$

where P_α is regularization term. However it is unclear how to devise such reguarlaization term in principle.

Based on AE, varational AE (VAE) derivied the loss function in a probablistic manner. We starts from $\log p(\boldsymbol{x})$:

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} && (1)\\
&= \log \int \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} && (2)\\
&\geq \int \log \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} && (3)\\
&= \mathbf{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z})] - \int \log \frac{p(\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} q(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} && (4)\\
&= \mathbf{E}_q[\log p(\boldsymbol{x}|\boldsymbol{z})] + \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) && (5)
\end{aligned}$$

The first term in the above equation is the log-likelihood of decoder output, while the second term minimize the KL divergence between encoder output and the target encoder distribution. Now the P_α in the Eq (1) has a probabilistic definition.

Denoise Diffusion Probablistic Model (DDPM)

Forward (diffusion) and Backward (denoise) process

In the DDPM, we start from $\boldsymbol{x}_0$, whose distribution is unknown. At each step t, a diffusion process is used:

$$\boldsymbol{x}_t = \alpha_t\, \boldsymbol{x}_{t-1} + \beta_t\, \boldsymbol{\varepsilon}_t$$

where $\boldsymbol{\varepsilon}_t$ draws from zero-mean unit-variance Gaussion distribution. Additionally, $\alpha_t^2 + \beta_t^2 = 1$. We have the following attributes regarding $\boldsymbol{x}_t$:

1. **Forward Process**: given $\boldsymbol{x}_{t-1}$, it is straightfoward to know that $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ follow a Gaussin distribution with $\alpha_t\, \boldsymbol{x}_{t-1}$ as mean and $\beta_t^2 \boldsymbol{I}$ as the variance.
2. **Fast Forward Process**: A nice property of DDPM is that the conditional distribution of $\boldsymbol{x}_t$ given $\boldsymbol{x}_0$ can be calculated explicitly without going through the recrusive process, i.e.,

$$
\begin{aligned}
\boldsymbol{x}_t &= \alpha_t \boldsymbol{x}_{t-1} + \beta \boldsymbol{\varepsilon_t} \\
&= \alpha_t \alpha_{t-1} ... \alpha_1 \boldsymbol{x}_0 + (\alpha_t ... \alpha_2)\beta_1 \boldsymbol{\varepsilon}_1 + ... + \beta_t \boldsymbol{\varepsilon}_t
\end{aligned} \tag{6}
$$

Except for the first term in Eq. (6), each term is a zero-mean, unit-variance Gaussion noise, therefore, Eq (6) can be also written as:

$$
\boldsymbol{x}_t = \overline{\alpha}_t \boldsymbol{x}_0 + \overline{\beta}_t \overline{\boldsymbol{\varepsilon}_t} \tag{7}
$$

where $\overline{\alpha}_t = \prod_{\tau=1}^{t} \alpha_\tau$, $\overline{\beta}_t = \sqrt{1 - \overline{\alpha}_t^2}$ and $\overline{\boldsymbol{\varepsilon}}_t$ is again a zero-mean, uni-variance Gaussin.

3. **Reverse Process**: However, we don't know the conditional distribution of $\boldsymbol{x}_{t-1}$ given $\boldsymbol{x}_t$. We only know that for small enough β_t, it is a still a Gaussian distribution. We use neural network (with parameter θ) to estimate the mean and variance, given $\boldsymbol{x}_t$ and t, i.e.,

$$
p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t))
$$

4. **Conditional Reverse Process**: Though we don't know the explicit form of reverse probability, a nice property of diffusion model is that $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ is a Gaussian distribution. This can be proved by the following deduction:

$$
p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{p(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{p(\boldsymbol{x}_t|\boldsymbol{x}_0)} = \frac{p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{p(\boldsymbol{x}_t|\boldsymbol{x}_0)}
$$

Since we know that $p(\boldsymbol{x}_\tau|\boldsymbol{x}_0)$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{x}_\tau; \overline{\alpha}_\tau \boldsymbol{x}_0, \overline{\beta}_\tau \boldsymbol{I})$ for $\tau > 1$ from **Forward Process**, we then have the following equations:

$$
p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \sim \exp(-\frac{1}{2\beta_t^2}(\boldsymbol{x}_t - \alpha_t \boldsymbol{x}_{t-1})^2) \cdot \exp(-\frac{1}{2\overline{\beta}_{t-1}^2}(\boldsymbol{x}_{t-1} - \overline{\alpha}_{t-1}\boldsymbol{x}_0)^2) \cdot \exp(\frac{1}{2\overline{\beta}_t^2}(\boldsymbol{x}_t - \overline{\alpha}_t \boldsymbol{x}_0)^2)
$$

.

Note that the above equation can be written as:

$$
\begin{aligned}
p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) &\sim \exp(-\frac{1}{2}(a\boldsymbol{x}_{t-1}^2 - 2b\boldsymbol{x}_{t-1} + c)) \\
&\sim \exp(-\frac{1}{2 \cdot 1/a}(\boldsymbol{x}_{t-1} - \frac{b}{a})^2)
\end{aligned}
$$

with (note that $\alpha_t^2 + \beta_t^2 = 1$, $\overline{\alpha}_t^2 + \overline{\beta}_t^2 = 1$ and $\overline{\alpha}_t = \overline{\alpha}_{t-1}\alpha_t$)

$$
\begin{aligned}
a &= \frac{\alpha_t^2}{\beta_t^2} + \frac{1}{\overline{\beta}_{t-1}^2} = \frac{\alpha_t^2(1-\overline{\alpha}_{t-1}^2)+\beta_t^2}{\beta_t^2(1-\overline{\alpha}_{t-1}^2)} = \frac{1-\overline{\alpha}_t^2}{\beta_t^2(1-\overline{\alpha}_{t-1}^2)} = \frac{\overline{\beta}_t^2}{\beta_t^2\overline{\beta}_{t-1}^2} = (\frac{\overline{\beta}_t}{\beta_t\overline{\beta}_{t-1}})^2 \\
b &= \frac{\alpha_t}{\beta_t^2}\boldsymbol{x}_t + \frac{\overline{\alpha}_{t-1}}{\overline{\beta}_{t-1}^2}\boldsymbol{x}_0 \\
\frac{b}{a} &= \frac{\alpha_t(1-\overline{\alpha}_{t-1}^2)}{1-\overline{\alpha}_t^2}\boldsymbol{x}_t + \frac{\beta_t^2\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t^2}\boldsymbol{x}_0 \\
&= \frac{\alpha_t\overline{\beta}_{t-1}^2}{\overline{\beta}_t^2}\boldsymbol{x}_t + \frac{\overline{\alpha}_{t-1}\beta_t^2}{\overline{\beta}_t^2}\boldsymbol{x}_0
\end{aligned}
$$

Therefore, we can see that $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ is again Gaussian with $\frac{b}{a}$ as its mean and $\sqrt{\frac{1}{a}}$ as its variance. In other word, we can re-parameterize $\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0$ by

$$
\boldsymbol{x}_{t-1} = \frac{\alpha_t\overline{\beta}_{t-1}^2}{\overline{\beta}_t^2}\boldsymbol{x}_t + \frac{\beta_t^2\overline{\alpha}_{t-1}}{\overline{\beta}_t^2}\boldsymbol{x}_0 + \frac{\beta_t\overline{\beta}_{t-1}}{\overline{\beta}_t}\boldsymbol{\varepsilon}
$$

with $\boldsymbol{\varepsilon}$ a sample from zero-mean, unit-variance Gaussian.
Using the **Fast forward** property, we know that

$$
\boldsymbol{x}_0 = \frac{1}{\overline{\alpha}_t}(\boldsymbol{x}_t - \overline{\beta}_t\overline{\boldsymbol{\varepsilon}_t})
$$

By combing the above 2 equations, we have:

$$
\boldsymbol{x}_{t-1} = \frac{1}{\alpha_t}(\boldsymbol{x}_t - \frac{1-\alpha_t^2}{\overline{\beta}_t}\overline{\boldsymbol{\epsilon}}_t) + \frac{\beta_t\overline{\beta}_{t-1}}{\overline{\beta}_t}\boldsymbol{\varepsilon}
$$

Variational EM to optimize θ

With the above properties, we can now derive the variational EM algorithm to maximize data distribution $p_\theta(\boldsymbol{x}_0)$ with respect to θ. Since only $\boldsymbol{x}_0$ is observed, and $\boldsymbol{x}_1, ...\boldsymbol{x}_T$ are latent, they can be treated as the latent variable $\boldsymbol{z}$ in Eq. (5). Using $q = p(\boldsymbol{x}_1, ...\boldsymbol{x}_T|\boldsymbol{x}_0)$ and follow Eq. (3), we can have the following:

$$
\log p_\theta(\boldsymbol{x}_0) \geq \mathbf{E}_q[\log \frac{p(\boldsymbol{x}_0,\cdots,\boldsymbol{x}_T)}{q(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T|\boldsymbol{x}_0)}] \tag{8}
$$

At the same time, we can use chain rule of probability to factorize $p(\boldsymbol{x}_0,\cdots,\boldsymbol{x}_T)$ in the following form:

$$
\begin{aligned}
p(\boldsymbol{x}_{0:T}) &= p(\boldsymbol{x}_T)p(\boldsymbol{x}_{0:T-1}|\boldsymbol{x}_T) && (9)\\
&= p(\boldsymbol{x}_T)p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)p(\boldsymbol{x}_{0:T-2}|\boldsymbol{x}_T,\boldsymbol{x}_{T-1}) && (10)\\
&= p(\boldsymbol{x}_T)p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T)p(\boldsymbol{x}_{0:T-2}|\boldsymbol{x}_{T-1}) && (11)\\
&= p(\boldsymbol{x}_T)\prod_{t=1}^{T} p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) && (12)
\end{aligned}
$$

Note that Eq (10) is possible because given $\boldsymbol{x}_{T-1}$, $\boldsymbol{x}_{0:T-2}$ is indepdent of $\boldsymbol{x}_T$.

Similarly, we have:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{13}$$

we also note that for $t > 1$, $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$ because $\boldsymbol{x}_t$ is conditional independent of $\boldsymbol{x}_0$ given $\boldsymbol{x}_{t-1}$. We can further reforulate $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \quad \forall t > 1$ by

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} \tag{14}$$

Cominging Eq. (12 - 14), we have:

$$\begin{aligned}
\log \frac{p(\boldsymbol{x}_0, \cdots, \boldsymbol{x}_T)}{q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\boldsymbol{x}_0)} &= \log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \\
&\quad - \log p(\boldsymbol{x}_1|\boldsymbol{x}_0) - \sum_{t=2}^{T} (\log q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) + \log q(\boldsymbol{x}_t|\boldsymbol{x}_0) - \log q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)) \\
&= \log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) - \log q(\boldsymbol{x}_T|\boldsymbol{x}_0) - \sum_{t=2}^{T} \log q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \\
&= \log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) + \sum_{t=2}^{T} \log \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)}
\end{aligned} \tag{15}$$

Therefore,

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}_0) &\geq \mathbf{E}_q[\log \frac{p(\boldsymbol{x}_0, \cdots, \boldsymbol{x}_T)}{q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|\boldsymbol{x}_0)}] \\
&= -\mathcal{D}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T)) - \sum_{t=2}^{T} \mathcal{D}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p(\boldsymbol{x}_{t-1}||\boldsymbol{x}_t)) + \mathrm{E}_q\left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]
\end{aligned} \tag{16}$$

where $\mathcal{D}(p||q)$ is the KL-diveragence between p and q. Let's denote:

- $L_0 = -\mathrm{E}_q[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)]$
- $L_{t-1} = \mathcal{D}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p(\boldsymbol{x}_{t-1}||\boldsymbol{x}_t)), \quad t = 2, \cdots, T$
- $L_T = \mathcal{D}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T))$

It is also noted that L_T is not a function of θ. Then the loss function of θ becomes:

$$\mathcal{L}(\theta) = L_0 + \sum_{t=1}^{T-1} L_t \tag{17}$$

Recall that:

- $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \frac{1}{\alpha_t}(\boldsymbol{x}_t - \frac{\beta_t^2}{\overline{\beta}_t}\overline{\boldsymbol{\varepsilon}}_t), \frac{\beta_t\overline{\beta}_{t-1}}{\overline{\beta}_t}\boldsymbol{I})$;
- $p(\boldsymbol{x}_{t-1}||\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t))$. For simplicity of derivation, we can reparameterize $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\alpha_t}(\boldsymbol{x}_t - \frac{\beta_t^2}{\overline{\beta}_t}\boldsymbol{\varepsilon}(\boldsymbol{x}_t, t))$.
- The KL distance between two Gaussian distributions has a closed form, i.e.,

$$\mathrm{KL}(\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) = \frac{1}{2}\left[\log\frac{\det(\boldsymbol{\Sigma}_2)}{\det\boldsymbol{\Sigma}_1} + \mathrm{tr}(\boldsymbol{\Sigma_2}^{-1}\boldsymbol{\Sigma_1}) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right] \tag{18}$$

Therefore, we have

$$L_t(\theta) = \frac{\beta_t^4}{2\alpha_t^2\overline{\beta}_t^2\det(\boldsymbol{\Sigma}_\theta)} \cdot \mathbf{E}_{\boldsymbol{x}_0,\boldsymbol{\varepsilon}_t}(\overline{\boldsymbol{\varepsilon}}_t - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t))^2 \tag{19}$$

$$= \frac{\beta_t^4}{2\alpha_t^2\overline{\beta}_t^2\det(\boldsymbol{\Sigma}_\theta)} \cdot \mathbf{E}_{\boldsymbol{x}_0,\boldsymbol{\varepsilon}_t}(\overline{\boldsymbol{\varepsilon}}_t - \boldsymbol{\varepsilon}_\theta(\overline{\alpha_t}\boldsymbol{x}_0 + \overline{\beta_t}\overline{\boldsymbol{\varepsilon}}_t, t))^2 \tag{20}$$

Note the above expectation is taken over $\boldsymbol{x}_0$ which is drawn from the data distribution and $\overline{\boldsymbol{\varepsilon}}_t$ which is a zero-mean, unit-variance Gaussian distribution. Emprically, it is found that training the following simplified loss function yield better results:

$$\mathcal{L}_{\mathrm{simple}} = \mathbf{E}_{t,\boldsymbol{x}_0,\overline{\varepsilon}_t}(\overline{\boldsymbol{\varepsilon}}_t - \boldsymbol{\varepsilon}_\theta(\overline{\alpha_t}\boldsymbol{x}_0 + \overline{\beta_t}\overline{\boldsymbol{\varepsilon}}_t, t))^2 \tag{21}$$

Based on this, the following training and inference algorithm can be derived:

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2$
6: **until** converged

Algorithm 2 Sampling

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

Interpretation of the Training and Sampling Process

In the previous section, we derive the training and sampling process in a mathematical rigorous way. On the other hand, it may not be easy to understand the algorithms. Here we provide a few approches to interpret how the training and sampling algorithm is derived in an intuitive manner.

Denoising perspective

The first method approaches to the problem from the denoising perspective. By the definition of reverse process, $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ is to recover the mean of $\boldsymbol{x}_{t-1}$, thefore a reasonable loss function to optimize is thus:

$$\mathcal{L}_{\mathrm{denoise}} = \mathbf{E}_{t,\boldsymbol{x}_{t-1},\boldsymbol{x}_t}\|\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)\|^2 \tag{22}$$

Since we don't know the distribution of $\boldsymbol{x}_t$ or $\boldsymbol{x}_{t-1}$ and their joint distribution, we cannot determine the above loss function. Instead, we know that (from the forward process):

$$\boldsymbol{x}_{t-1} = \frac{1}{\alpha_t}(\boldsymbol{x}_t - \beta_t \boldsymbol{\varepsilon}_t) \tag{23}$$

Accordingly, we re-parameterize $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ as

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) = \frac{1}{\alpha_t}(\boldsymbol{x}_t - \beta_t \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)) \tag{24}$$

Then the loss function becomes

$$\mathcal{L}_{\text{denoise}} = \mathbf{E}_{t,\boldsymbol{x}_t} \frac{\beta_t^2}{\alpha_t^2} \|\boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)\|^2 \tag{25}$$

To sample $\boldsymbol{x}_t$, recall that (by the fast-forward process),

$$\begin{align} \boldsymbol{x}_t &= \overline{\alpha}_t \boldsymbol{x}_0 + \overline{\beta}_t \overline{\boldsymbol{\varepsilon}}_t \tag{26} \\ &= \alpha_t(\overline{\alpha}_{t-1}\boldsymbol{x}_0 + \overline{\beta}_{t-1}\overline{\boldsymbol{\varepsilon}}_{t-1}) + \beta_t \boldsymbol{\varepsilon}_t \tag{27} \end{align}$$

Plugging Eq. (27) into the loss function Eq. (22), (note that we cannot plug Eq. (26) into Eq. (22), because $\overline{\boldsymbol{\varepsilon}}_t$ is a function of $\boldsymbol{\varepsilon}_t$, so they cannot be sampled independently), we got

$$\mathcal{L}_{\text{denoise}} = \mathbf{E}_{t,\boldsymbol{\varepsilon}_t,\overline{\boldsymbol{\varepsilon}}_{t-1}} \frac{\beta_t^2}{\alpha_t^2} \|\boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_\theta(\overline{\alpha}_t \boldsymbol{x}_0 + \alpha_t \overline{\beta}_{t-1}\overline{\boldsymbol{\varepsilon}}_{t-1} + \beta_t \boldsymbol{\varepsilon}_t, t)\|^2 \tag{28}$$

We further noting that:

$$\overline{\beta}_t \overline{\boldsymbol{\varepsilon}}_t = \alpha_t \overline{\beta}_{t-1}\overline{\boldsymbol{\varepsilon}}_{t-1} + \beta_t \boldsymbol{\varepsilon}_t \tag{29}$$

is independent of $\overline{\beta}_t \boldsymbol{w} = \beta_t \overline{\boldsymbol{\varepsilon}}_{t-1} - \alpha_t \overline{\beta}_{t-1} \boldsymbol{\varepsilon}_t$ and

$$\boldsymbol{\varepsilon}_t = \frac{\beta_t \overline{\boldsymbol{\varepsilon}}_t - \alpha_t \overline{\beta}_{t-1} \boldsymbol{w}}{\overline{\beta}_t} \tag{30}$$

With the above changing of variable, the loss function can then be written as:

$$\mathcal{L}_{\text{denoise}} = \mathbf{E}_{t,\overline{\boldsymbol{\varepsilon}}_t} \frac{\beta_t^2}{\alpha_t^2} \|\frac{\beta_t}{\overline{\beta}_t}\overline{\boldsymbol{\varepsilon}}_t - \boldsymbol{\varepsilon}_\theta(\overline{\alpha}_t \boldsymbol{x}_0 + \overline{\beta}_t \overline{\boldsymbol{\varepsilon}}_t, t)\|^2 + C \tag{31}$$

where C is a constant (only related to $\boldsymbol{w}$). Also note that after the changing of variables, $\boldsymbol{\varepsilon}_\theta(\cdot, \cdot)$ in Eq (28) and (31) are different functions (one is trying to denoise one step noise $\boldsymbol{\varepsilon}_t$, and the other is trying to denoise cumulative noise $\overline{\boldsymbol{\varepsilon}}_t$)