Gaussian Diffusion

Lan Zhang

1 Basic Assumption

The data distribution is gradually converted into a well behaved (analytically tractable) distribution $\pi(\mathbf{y})$ by repeated application of a Markov diffusion kernel $T_\pi(\mathbf{y}|\mathbf{y}';\beta)$ for $\pi(\mathbf{y})$, where β is the diffusion rate.

$$\pi(\mathbf{y}) = \int T_\pi(\mathbf{y}|\mathbf{y}';\beta)\pi(\mathbf{y}')\, d\mathbf{y}'$$

We model the conditional probability under this Markov chain at timestamp t as an isotropic Gaussian:

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(0\cdots(t-1))}) = q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = T_\pi(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)};\beta_t) = \mathcal{N}(\mathbf{x}^{(t)};\sqrt{1-\beta_t}\mathbf{x}^{(t-1)},\beta_t\mathbf{I}),$$

and the joint distribution across time T is as:

$$q(\mathbf{x}^{(0\cdots T)}) = q(\mathbf{x}^{(0)})\prod_{t=1}^{T} q(\mathbf{x}^{(t)}|\mathbf{x}^{(0\cdots(t-1))}) = q(\mathbf{x}^{(0)})\prod_{t=1}^{T} q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}).$$

For continuous Gaussian diffusion (limit of small step size β_t), the reversal of the diffusion process has the identical functional form as the forward process[1]. Therefore, if β_t is small, $q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ should also be a Gaussian. Similar to Variational Autoencoder (VAE), we use $p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t-1)}; f_{\boldsymbol{\mu}}^{(t)}(\mathbf{x}^{(t)}), f_{\boldsymbol{\Sigma}}^{(t)}(\mathbf{x}^{(t)}))$ to recognize this distribution in the reverse Markov process. We also assume that at the end of the diffusion process $p(\mathbf{x}^{(T)}) = \pi(\mathbf{x}^{(T)}) = \mathcal{N}(\mathbf{x}^{(T)};\mathbf{0},\mathbf{I})$. Then the probability of the original data $\mathbf{x}^{(0)}$ under this process is:

$$\begin{aligned} p(\mathbf{x}^{(0)}) &= \int p(\mathbf{x}^{(0\cdots T)})\, d\mathbf{x}^{(1\cdots T)} \\ &= \int p(\mathbf{x}^{(T)})\prod_{t=1}^{T} p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t\cdots T)})\, d\mathbf{x}^{(1\cdots T)} \\ &= \int p(\mathbf{x}^{(T)})\prod_{t=1}^{T} p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})\, d\mathbf{x}^{(1\cdots T)}. \end{aligned}$$

[1] On the theory of stochastic processes, with particular reference to applications, W. Feller, 1949

2 Objective Function

The common objective function in data modeling, cross entropy loss becomes:

$$\begin{aligned} L &= \int q(\mathbf{x}^{(0)}) \log p(\mathbf{x}^{(0)}) \, d\mathbf{x}^{(0)} \\ &= \int q(\mathbf{x}^{(0)}) \log(\int p(\mathbf{x}^{(T)}) \prod_{t=1}^{T} p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) \, d\mathbf{x}^{(1\cdots T)}) \, d\mathbf{x}^{(0)}. \end{aligned}$$

In this form, the link between p and q is "weak" as there is no measure about how well p and q complement each other at time stemp t. Recall that q also has an similar product expression, hence we can rewrite:

$$\begin{aligned} L &= \int q(\mathbf{x}^{(0)}) \log(\int p(\mathbf{x}^{(T)}) \prod_{t=1}^{T} p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) \, d\mathbf{x}^{(1\cdots T)}) \, d\mathbf{x}^{(0)} \\ &= \int q(\mathbf{x}^{(0)}) \log(\int p(\mathbf{x}^{(T)}) \frac{q(\mathbf{x}^{(0\cdots T)})}{q(\mathbf{x}^{(0)})} \prod_{t=1}^{T} \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \, d\mathbf{x}^{(1\cdots T)}) \, d\mathbf{x}^{(0)}. \end{aligned}$$

There is a multiple integral in the $\log$ function which makes it hard to calcuate. Note that

$$\frac{q(\mathbf{x}^{(0\cdots T)})}{q(\mathbf{x}^{(0)})} = q(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)})$$

is a probability density function in the internal multiple integral, which means that

$$\int q(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}) \, d\mathbf{x}^{(1\cdots T)} = 1.$$

Using Jensen's inequality, instead of directly optimizing L, we optimize the lower bound K of it:

$$\begin{aligned} L &= \int q(\mathbf{x}^{(0)}) \log(\int p(\mathbf{x}^{(T)}) \frac{q(\mathbf{x}^{(0\cdots T)})}{q(\mathbf{x}^{(0)})} \prod_{t=1}^{T} \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \, d\mathbf{x}^{(1\cdots T)}) \, d\mathbf{x}^{(0)} \\ &\geq \int q(\mathbf{x}^{(0)}) (\int \frac{q(\mathbf{x}^{(0\cdots T)})}{q(\mathbf{x}^{(0)})} \log(p(\mathbf{x}^{(T)}) \prod_{t=1}^{T} \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})}) \, d\mathbf{x}^{(1\cdots T)}) \, d\mathbf{x}^{(0)} = K. \end{aligned}$$

Combining the integral expression and changing the product to sum, we have:

$$\begin{aligned} K &= \int q(\mathbf{x}^{(0\cdots T)}) \sum_{t=1}^{T} \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \, d\mathbf{x}^{(0\cdots T)} + \int q(\mathbf{x}^{(0\cdots T)}) \log p(\mathbf{x}^{(T)}) \, d\mathbf{x}^{(0\cdots T)} \\ &= \sum_{t=1}^{T} \int q(\mathbf{x}^{(0\cdots T)}) \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \, d\mathbf{x}^{(0\cdots T)} + \int q(\mathbf{x}^{(0\cdots T)}) \log p(\mathbf{x}^{(T)}) \, d\mathbf{x}^{(0\cdots T)}. \end{aligned}$$

In each term, some variables of the multiple integral do not influence the calculation. We remove these variables except $\mathbf{x}^{(0)}$ to keep the connection between intermediate codes and the original data. This leads to:

$$\begin{aligned} K = &\sum_{t=2}^{T} \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t-1)} \, d\mathbf{x}^{(t)} \\ &+ \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) \log \frac{p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})}{q(\mathbf{x}^{(1)}|\mathbf{x}^{(0)})} \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(1)} + \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(T)}) \log p(\mathbf{x}^{(T)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(T)}. \end{aligned}$$

Each term in the sum can be rewrited as:

$$\begin{aligned}
&\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t-1)} \, d\mathbf{x}^{(t)} \\
=&\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) (\int q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \, d\mathbf{x}^{(t-1)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t)} \\
=&\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) f(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t)}.
\end{aligned}$$

Expand $f(\mathbf{x}^{(0)}, \mathbf{x}^{(t)})$, we have:

$$\begin{aligned}
f(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) &= \int q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})} \, d\mathbf{x}^{(t-1)} \\
&= \int q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \mathbf{x}^{(0)})} \, d\mathbf{x}^{(t-1)} \\
&= \int q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)})} \, d\mathbf{x}^{(t-1)}.
\end{aligned}$$

Hence, the integral:

$$\begin{aligned}
&\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) f(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t)} \\
=&\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) (\int q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \log \frac{p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)})}{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)})} \, d\mathbf{x}^{(t-1)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t)} \\
=&-\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \| p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t)} \\
&+\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) \log \frac{q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)})}{q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)})} \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t-1)} \, d\mathbf{x}^{(t)}
\end{aligned}$$

Note that:

$$\begin{aligned}
&\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) \log \frac{q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)})}{q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)})} \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t-1)} \, d\mathbf{x}^{(t)} \\
=&\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t)}) \log \frac{q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})} \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t-1)} \, d\mathbf{x}^{(t)} \\
=&\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)}) \log q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t-1)} - \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \log q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t)}.
\end{aligned}$$

The expression of K becomes:

$$\begin{aligned}
K =& -\sum_{t=2}^{T} \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \| p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t)} \\
&+\sum_{t=2}^{T} (\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)}) \log q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t-1)} - \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \log q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(t)}) \\
&+\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) \log \frac{p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})}{q(\mathbf{x}^{(1)}|\mathbf{x}^{(0)})} \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(1)} + \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(T)}) \log p(\mathbf{x}^{(T)}) \, d\mathbf{x}^{(0)} \, d\mathbf{x}^{(T)}.
\end{aligned}$$

With further simplification, the final objective function is:

$$K = -\sum_{t=2}^{T} \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) \| p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}))\, d\mathbf{x}^{(0)}\, d\mathbf{x}^{(t)}$$
$$+ \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(1)}) \log p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})\, d\mathbf{x}^{(0)}\, d\mathbf{x}^{(1)} + \int q(\mathbf{x}^{(0)}, \mathbf{x}^{(T)}) \log \frac{p(\mathbf{x}^{(T)})}{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}\, d\mathbf{x}^{(0)}\, d\mathbf{x}^{(T)}.$$

The KL-divergence is a function of $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(t)}$ and can be computed analytically. Intuitively, this objective function is a combination of a series of sub-functions and each sub-function is only relevant to one timestamp.

2.1 Conditional Distribution $q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})$

Recall that $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{1-\beta_t}\mathbf{x}^{(t-1)}, \beta_t\mathbf{I})$. Using the reparameterization trick, in the forward process, we have:

$$\mathbf{x}^{(t)} = \sqrt{1-\beta_t}\mathbf{x}^{(t-1)} + \sqrt{\beta_t}\boldsymbol{\epsilon}^{(t-1)} = \sqrt{1-\beta_t}(\sqrt{1-\beta_{t-1}}\mathbf{x}^{(t-2)} + \sqrt{\beta_{t-1}}\boldsymbol{\epsilon}^{(t-2)}) + \sqrt{\beta_t}\boldsymbol{\epsilon}^{(t-1)}$$

where $\boldsymbol{\epsilon}^{(t-1)}, \boldsymbol{\epsilon}^{(t-2)} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \mathbf{I})$. Because variables of $\boldsymbol{\epsilon}$ are independent and share the same variance, according to sum of normally distributed random variables, we have

$$\sqrt{1-\beta_t}\sqrt{\beta_{t-1}}\boldsymbol{\epsilon}^{(t-2)} + \sqrt{\beta_t}\boldsymbol{\epsilon}^{(t-1)} \sim \mathcal{N}(\boldsymbol{\epsilon}; (\mathbf{0}, \beta_t + \beta_{t-1} - \beta_t\beta_{t-1})\mathbf{I}) = \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, (1-(1-\beta_t)(1-\beta_{t-1}))\mathbf{I}).$$

For better notations, let $\alpha_t = 1 - \beta_t$. We have $\mathbf{x}^{(t)} \sim \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}^{(t-2)}, (1-\alpha_t\alpha_{t-1})\mathbf{I})$. Repeat this deduction $t-1$ times, we have:

$$\mathbf{x}^{(t)} \sim \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\prod_{i=1}^{t}\alpha_i}\mathbf{x}^{(0)}, (1-\prod_{i=1}^{t}\alpha_i)\mathbf{I}).$$

This is equivalent to:

$$q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\prod_{i=1}^{t}\alpha_i}\mathbf{x}^{(0)}, (1-\prod_{i=1}^{t}\alpha_i)\mathbf{I}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\bar{\alpha}_t}\mathbf{x}^{(0)}, (1-\bar{\alpha}_t)\mathbf{I}).$$

2.2 Computing KL-divergence

The KL-divergence between two multivariate Gaussian (normal) distributions is:

$$D_{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \| \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) = \frac{1}{2}(\log\frac{\det\boldsymbol{\Sigma}_2}{\det\boldsymbol{\Sigma}_1} - n + \mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)).$$

The third term in K can be converted to:

$$\int q(\mathbf{x}^{(0)}, \mathbf{x}^{(T)}) \log \frac{p(\mathbf{x}^{(T)})}{q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)})}\, d\mathbf{x}^{(0)}\, d\mathbf{x}^{(T)} = -\int q(\mathbf{x}^{(0)}) D_{KL}(q(\mathbf{x}^{(T)}|\mathbf{x}^{(0)}) \| p(\mathbf{x}^{(T)}))\, d\mathbf{x}^{(0)}$$
$$= -\int \frac{1}{2}(\bar{\alpha}_T \|\mathbf{x}^{(0)}\|^2 - n\bar{\alpha}_T - n\log(1-\bar{\alpha}_T)) q(\mathbf{x}^{(0)})\, d\mathbf{x}^{(0)}.$$

We already know $p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}) = \mathcal{N}(\mathbf{x}^{(t-1)}; f_{\boldsymbol{\mu}}^{(t)}(\mathbf{x}^{(t)}), f_{\boldsymbol{\Sigma}}^{(t)}(\mathbf{x}^{(t)}))$, and for $t > 1$:

$$q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)}) = \frac{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}, \mathbf{x}^{(0)})q(\mathbf{x}^{(0)}, \mathbf{x}^{(t-1)})}{q(\mathbf{x}^{(0)}, \mathbf{x}^{(t)})} = \frac{q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)})}{q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})}$$

$$= \sqrt{(\frac{1-\bar{\alpha}_t}{2\pi(1-\alpha_t)(1-\bar{\alpha}_{t-1})})^n} \exp(-(\frac{(\mathbf{x}^{(t)} - \sqrt{\alpha_t}\mathbf{x}^{(t-1)})^2}{2(1-\alpha_t)} + \frac{(\mathbf{x}^{(t-1)} - \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}^{(0)})^2}{2(1-\bar{\alpha}_{t-1})} - \frac{(\mathbf{x}^{(t)} - \sqrt{\bar{\alpha}_t}\mathbf{x}^{(0)})^2}{2(1-\bar{\alpha}_t)}))$$

$$= \mathcal{N}(\mathbf{x}^{(t-1)}; \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}\mathbf{x}^{(t)} + (1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}^{(0)}}{1-\bar{\alpha}_t}, \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}).$$

Therefore, $D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})\|p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}))$ can be analytically computed.

2.3 Monte Carlo Estimates

In practice, we treat each datapoint $\mathbf{x}^{(0)}$ as an individual sample from the underlying distribution $q(\mathbf{x}^{(0)})$. Hence the objective function can be estimated as:

$$K \simeq -\sum_{t=2}^{T} \int q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})\|p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}))\, d\mathbf{x}^{(t)}$$
$$+ \int q(\mathbf{x}^{(1)}|\mathbf{x}^{(0)}) \log p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)})\, d\mathbf{x}^{(1)} - \frac{1}{2}(\bar{\alpha}_T\|\mathbf{x}^{(0)}\|^2 - n\bar{\alpha}_T - n\log(1-\bar{\alpha}_T)).$$

The integral here is still not friendly and cannot be computed directly. Therefore, we use Monte Carlo estimates again to avoid integral calculation and yield:

$$K \simeq -\sum_{t=2}^{T} D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})\|p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})) + \log p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}) - \frac{1}{2}(\bar{\alpha}_T\|\mathbf{x}^{(0)}\|^2 - n\bar{\alpha}_T - n\log(1-\bar{\alpha}_T)).$$

In this estimation, $\mathbf{x}^{(0)}$ is given from the dataset and for each time step t, $\mathbf{x}^{(t)}$ is a sample from $q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\bar{\alpha}_t}\mathbf{x}^{(0)}, (1-\bar{\alpha}_t)\mathbf{I})$.

3 Optimization

Use notations:

$$\boldsymbol{\mu}_t = f_{\boldsymbol{\mu}}^{(t)}(\mathbf{x}^{(t)}), \boldsymbol{\Sigma}_t = f_{\boldsymbol{\Sigma}}^{(t)}(\mathbf{x}^{(t)}).$$

Let:

$$\mathcal{L}(\mathbf{x}^{(0\cdots T)}; \alpha_{1\cdots T}, \boldsymbol{\mu}_{1\cdots T}, \boldsymbol{\Sigma}_{1\cdots T}) = \sum_{t=2}^{T} D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})\|p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}))$$
$$- \log p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}) + \frac{1}{2}(\bar{\alpha}_T\|\mathbf{x}^{(0)}\|^2 - n\bar{\alpha}_T - n\log(1-\bar{\alpha}_T)).$$

The optimization problem is:

$$\operatorname{argmax} K = \operatorname*{argmin}_{\alpha_{1\cdots T},\, \boldsymbol{\mu}_{1\cdots T},\, \boldsymbol{\Sigma}_{1\cdots T}} \mathcal{L}(\mathbf{x}^{(0\cdots T)}; \alpha_{1\cdots T}, \boldsymbol{\mu}_{1\cdots T}, \boldsymbol{\Sigma}_{1\cdots T}).$$

3.1 Choice of $\mathbf{\Sigma}_t$

The analytical computation of KL-divergence requires the computation of determinant and inverse of $\mathbf{\Sigma}_t$, which can be time-consuming. The common practice is to set it to a diagonal matrix:

$$\mathbf{\Sigma}_t = \text{diag}(\sigma_{t,1}^2, \sigma_{t,2}^2, \cdots, \sigma_{t,n}^2).$$

However, this setting has two potential shortcomings. Firstly, we need to estimate nT variances to obtain $\mathbf{\Sigma}_{1\cdots T}$. For data with high dimensionality, the estimation of variances would have large variance. Secondly, this setting indicates that the diffusion in the reverse process is anisotropic. From the physical perspective, this can lead to unwanted directional bias in the diffusion, and the model might translate this into bias from the data.

To alleviate these two drawbacks, we choose $\mathbf{\Sigma}_t = \sigma_t^2\mathbf{I}$, and for $t > 1$ the KL-divergence becomes:

$$D_{KL}(q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})\|p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})) = \frac{1}{2}(n\log\frac{\sigma_t^2}{\bar{\sigma}_t^2} - n + n\frac{\bar{\sigma}_t^2}{\sigma_t^2} + \frac{1}{\sigma_t^2}\|\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_t\|^2),$$

where

$$\bar{\boldsymbol{\mu}}_t = \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}\mathbf{x}^{(t)} + (1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}^{(0)}}{1-\bar{\alpha}_t}, \bar{\sigma}_t^2 = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}.$$

In addition, define $\bar{\boldsymbol{\mu}}_1 = \mathbf{x}^{(0)}$. We have:

$$\log p(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}) = -\frac{1}{2}(n\log 2\pi\sigma_1^2 + \frac{1}{\sigma_1^2}\|\boldsymbol{\mu}_1 - \mathbf{x}^{(0)}\|^2) = -\frac{1}{2}(n\log 2\pi\sigma_1^2 + \frac{1}{\sigma_1^2}\|\boldsymbol{\mu}_1 - \bar{\boldsymbol{\mu}}_1\|^2),$$

and

$$\mathcal{L} = \frac{1}{2}(\sum_{t=1}^{T}\frac{1}{\sigma_t^2}\|\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_t\|^2 + \bar{\alpha}_T\|\mathbf{x}^{(0)}\|^2 + \sum_{t=2}^{T}(n\log\frac{\sigma_t^2}{\bar{\sigma}_t^2} - n + n\frac{\bar{\sigma}_t^2}{\sigma_t^2}) + n\log 2\pi\sigma_1^2 - n\bar{\alpha}_T - n\log(1-\bar{\alpha}_T)).$$

Because the optimization is irrelevant to a constant scaling, we define the loss function as:

$$\mathcal{L}_0 = \sum_{t=1}^{T}\frac{1}{\sigma_t^2}\|\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_t\|^2 + \bar{\alpha}_T\|\mathbf{x}^{(0)}\|^2 + n(\sum_{t=2}^{T}(\log\frac{\sigma_t^2}{\bar{\sigma}_t^2} - 1 + \frac{\bar{\sigma}_t^2}{\sigma_t^2}) + \log 2\pi\sigma_1^2 - \bar{\alpha}_T - \log(1-\bar{\alpha}_T)).$$

3.2 Learnable Parameters Setting

In $\mathcal{L}_0$, learnable parameters are:

$$\alpha_{1\cdots T}, \boldsymbol{\mu}_{1\cdots T}, \sigma_{1\cdots T}^2.$$

We can consider $\alpha_{1\cdots T}$ as hyperparameters and set them to constants instead of learnable parameters during training. Then the optimization becomes:

$$\underset{\boldsymbol{\mu}_{1\cdots T},\, \sigma_{1\cdots T}^2}{\text{argmin}} \ \mathcal{L}_1, \text{ where } \mathcal{L}_1 = \sum_{t=1}^{T}\frac{1}{\sigma_t^2}\|\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_t\|^2 + n\sum_{t=1}^{T}\log\sigma_t^2 + n\sum_{t=2}^{T}\frac{\bar{\sigma}_t^2}{\sigma_t^2}.$$

In addition, to further simplify the optimization, we can set $\sigma_{1\cdots T}^2$ to constants. In this case, the optimization is to find $\boldsymbol{\mu}_{1\cdots T}$ which minimize:

$$\mathcal{L}_2 = \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \|\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_t\|^2.$$

The constant values of $\sigma_{1\cdots T}^2$ can be set as $\sigma_t^2 = \beta_t$ or $\sigma_t^2 = \bar{\sigma}_t^2$ for $t > 1$, which corresponds to the setting that $p(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)})$ has the same covariance matrix as $q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$ or $q(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}, \mathbf{x}^{(t)})$, respectively.

For $\boldsymbol{\mu}_{1\cdots T}$, we can use a shared deep neural network parameterized by θ to estimate them in three settings. The network takes a vector with the dimensionality of $\mathbf{x}^{(0)}$ as input and outputs a vector with the same dimensionality. In order to behave differently across different time steps, the network should also be time-aware.

3.2.1 Direct Prediction

$$\boldsymbol{\mu}_t = \hat{\boldsymbol{\mu}}_\theta(\mathbf{x}^{(t)}, t),$$

$$\mathcal{L}_2^{\text{dir}} = \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \|\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_t\|^2 = \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \|\hat{\boldsymbol{\mu}}_\theta(\mathbf{x}^{(t)}, t) - \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}\mathbf{x}^{(t)} + (1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}^{(0)}}{1-\bar{\alpha}_t}\|^2.$$

3.2.2 Reconstruction

$$\boldsymbol{\mu}_t = \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}\mathbf{x}^{(t)} + (1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_\theta^{(0)}(\mathbf{x}^{(t)}, t)}{1-\bar{\alpha}_t},$$

$$\mathcal{L}_2^{\text{rec}} = \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \|\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_t\|^2 = \sum_{t=1}^{T} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{\sigma_t^2(1-\bar{\alpha}_t)^2} \|\hat{\mathbf{x}}_\theta^{(0)}(\mathbf{x}^{(t)}, t) - \mathbf{x}^{(0)}\|^2.$$

3.2.3 Denoising

Recall that $\mathbf{x}^{(t)}$ is sampled from $q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}) = \mathcal{N}(\mathbf{x}^{(t)}; \sqrt{\bar{\alpha}_t}\mathbf{x}^{(0)}, (1-\bar{\alpha}_t)\mathbf{I})$ by using reparameterization trick $\mathbf{x}^{(t)} = \sqrt{\bar{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We have:

$$\mathbf{x}^{(0)} = \frac{\mathbf{x}^{(t)} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t}{\sqrt{\bar{\alpha}_t}},$$

$$\bar{\boldsymbol{\mu}}_t = \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}\mathbf{x}^{(t)} + (1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}^{(0)}}{1-\bar{\alpha}_t} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}^{(t)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t).$$

We set:

$$\boldsymbol{\mu}_t = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}^{(t)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}^{(t)}, t)).$$

Then the loss function becomes:

$$\mathcal{L}_2^{\text{den}} = \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \|\boldsymbol{\mu}_t - \bar{\boldsymbol{\mu}}_t\|^2 = \sum_{t=1}^{T} \frac{(1-\alpha_t)^2}{\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \|\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}^{(t)}, t) - \boldsymbol{\epsilon}_t\|^2.$$

We prefer the last two loss functions because they remove the simple linear transformation from the input to the output. Now, the optimization becomes $\operatorname{argmin}_\theta \mathcal{L}_2$.

3.3 Further Simplification for Training

In deep learning, to optimize $\mathcal{L}_2$ for one datapoint, we need to do sampling and forward the neural network for T times, and then do backward. For large T, the model needs to wait for a huge amount of time and use a large amount of memory for one update of parameters. This is both time-inefficient and memory-consuming. Since $\mathcal{L}_2$ is a sum of T sub-functions, like stochastic gradient descent to gradient descent, for each iteration, we can sample one t uniformly from $1 \cdots T$ and do optimization:

$$\operatorname*{argmin}_\theta \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{\sigma_t^2(1-\bar{\alpha}_t)^2} \|\hat{\mathbf{x}}_\theta^{(0)}(\mathbf{x}^{(t)}, t) - \mathbf{x}^{(0)}\|^2 \text{ or } \operatorname*{argmin}_\theta \frac{(1-\alpha_t)^2}{\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \|\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}^{(t)}, t) - \boldsymbol{\epsilon}_t\|^2.$$

This is equivalent to:

$$\operatorname*{argmin}_\theta \|\hat{\mathbf{x}}_\theta^{(0)}(\mathbf{x}^{(t)}, t) - \mathbf{x}^{(0)}\|^2 \text{ or } \operatorname*{argmin}_\theta \|\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}^{(t)}, t) - \boldsymbol{\epsilon}_t\|^2.$$

To keep the consistency, we define:

$$\mathcal{L}_{\text{simple}}^{\text{rec}} = \mathbb{E}_t(\|\hat{\mathbf{x}}_\theta^{(0)}(\mathbf{x}^{(t)}, t) - \mathbf{x}^{(0)}\|^2),$$

and

$$\mathcal{L}_{\text{simple}}^{\text{den}} = \mathbb{E}_t(\|\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}^{(t)}, t) - \boldsymbol{\epsilon}_t\|^2).$$

These two functions are used for the simplification of optimization.