List of Appendix

A. RayFlow

A.1. Flow Trajectory

Proposition 1. *Given data $\boldsymbol{x}_0$, pretrained mean $\boldsymbol{\epsilon}_\mu \in \mathbb{R}^d$ and variance $\sigma \in \mathbb{R}_{>0}$, and target diffusion $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{\epsilon}_\mu, \sigma^2 \boldsymbol{I})$, we can describe the diffusion process with the following Markov chain.*

Flow Trajectories. *Define probability flow path.*

$$\psi_t(\cdot|\boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + (1 - \sqrt{\alpha_t})\boldsymbol{\epsilon}_\mu + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon} \tag{24}$$

Forward Process. *Add noise to image data.*

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\epsilon}_\mu) = \mathcal{N}\left(\alpha_t \boldsymbol{x}_{t-1} + (1 - \alpha_t)\boldsymbol{\epsilon}_\mu, \beta_t^2 \sigma^2 \boldsymbol{I}\right)$$

Backward Process. *Denoise from image data.*

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{\epsilon}_\mu) = \mathcal{N}\left(\frac{1}{\alpha_t}\boldsymbol{x}_t - \frac{1 - \alpha_t}{\alpha_t}\boldsymbol{\epsilon}_\mu, \tilde{\beta}_t \sigma^2 \boldsymbol{I}\right) \tag{25}$$

$$\tilde{\beta}_t = \left(\frac{(1 - \alpha_t^2)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\right) \tag{26}$$

Proof. We split the proof process into two parts: **Forward Process** and **Backward Process**.

Forward Process

We set $\alpha_t^2 + \beta_t^2 = 1$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_t^2$, According to the forward process, we have:

$$\begin{aligned}
\boldsymbol{x}_t &= \alpha_t \boldsymbol{x}_{t-1} + (1 - \alpha_t)\boldsymbol{\epsilon}_\mu + \beta_t \boldsymbol{\epsilon}_t \\
&= \alpha_t(\alpha_{t-1}\boldsymbol{x}_{t-2} + (1 - \alpha_{t-1})\boldsymbol{\epsilon}_\mu + \beta_{t-1}\boldsymbol{\epsilon}_{t-1}) + (1 - \alpha_t)\boldsymbol{\epsilon}_\mu + \beta_t \boldsymbol{\epsilon}_t \\
&= \alpha_t \alpha_{t-1} \cdots \alpha_1 \boldsymbol{x}_0 \\
&\quad + \alpha_t \alpha_{t-1} \cdots \alpha_2 (1 - \alpha_1)\boldsymbol{\epsilon}_\mu + \cdots + \alpha_t \alpha_{t-1}(1 - \alpha_{t-2})\boldsymbol{\epsilon}_\mu + \alpha_t(1 - \alpha_{t-1})\boldsymbol{\epsilon}_\mu + (1 - \alpha_t)\boldsymbol{\epsilon}_\mu \\
&\quad + \alpha_t \alpha_{t-1} \cdots \alpha_2 \beta_1 \boldsymbol{\epsilon}_1 + \alpha_t \alpha_{t-1} \cdots \alpha_3 \beta_2 \boldsymbol{\epsilon}_2 + \cdots + \alpha_t \beta_{t-1}\boldsymbol{\epsilon}_t + \beta_t \boldsymbol{\epsilon}_t \\
&= \alpha_t \alpha_{t-1} \cdots \alpha_1 \boldsymbol{x}_0 + (1 - \alpha_t \alpha_{t-1} \cdots \alpha_1)\boldsymbol{\epsilon}_\mu \\
&\quad + \alpha_t \alpha_{t-1} \cdots \alpha_2 \beta_1 \boldsymbol{\epsilon}_1 + \alpha_t \alpha_{t-1} \cdots \alpha_3 \beta_2 \boldsymbol{\epsilon}_2 + \cdots + \alpha_t \beta_{t-1}\boldsymbol{\epsilon}_t + \beta_t \boldsymbol{\epsilon}_t
\end{aligned} \tag{27}$$

where

$$\begin{aligned}
&\alpha_t \alpha_{t-1} \cdots \alpha_2 \beta_1 \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\boldsymbol{0}, \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_2^2 \beta_1^2 \sigma^2 \boldsymbol{I}) \\
&\alpha_t \alpha_{t-1} \cdots \alpha_3 \beta_2 \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\boldsymbol{0}, \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_3^2 \beta_2^2 \sigma^2 \boldsymbol{I}) \\
&\cdots
\end{aligned} \tag{28}$$

we set $\alpha_t^2 + \beta_t^2 = 1$, then we have:

$$\begin{aligned}
&\alpha_t \alpha_{t-1} \cdots \alpha_2 \beta_1 \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\boldsymbol{0}, \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_2^2 (1 - \alpha_1^2) \sigma^2 \boldsymbol{I}) \\
&\alpha_t \alpha_{t-1} \cdots \alpha_3 \beta_2 \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\boldsymbol{0}, \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_3^2 (1 - \alpha_2^2) \sigma^2 \boldsymbol{I}) \\
&\cdots
\end{aligned} \tag{29}$$

based on the properties of Gaussian distributions:

$$\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu_1}, \sigma_1^2 \boldsymbol{I}), \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{\mu_2}, \sigma_2^2 \boldsymbol{I}), \tag{30}$$

$$\boldsymbol{a} + \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}, (\sigma_1^2 + \sigma_2^2)\boldsymbol{I}), \tag{31}$$

then

$$\begin{aligned}
&\alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_2^2 (1 - \alpha_1^2)\sigma^2 + \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_3^2 (1 - \alpha_2^2)\sigma^2 \\
&= \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_2^2 \sigma^2 - \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_1^2 \sigma^2 + \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_3^2 \sigma^2 - \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_2^2 \sigma^2 \\
&= \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_3^2 \sigma^2 - \alpha_t^2 \alpha_{t-1}^2 \cdots \alpha_1^2 \sigma^2.
\end{aligned} \tag{32}$$

$$\begin{aligned}&\alpha_t\alpha_{t-1}\cdots\alpha_2\beta_1\boldsymbol{\epsilon}_1+\alpha_t\alpha_{t-1}\cdots\alpha_3\beta_2\boldsymbol{\epsilon}_2\\&\sim\mathcal{N}(\mathbf{0},(\alpha_t^2\alpha_{t-1}^2\cdots\alpha_3^2-\alpha_t^2\alpha_{t-1}^2\cdots\alpha_1^2)\sigma^2\boldsymbol{I}).\end{aligned}\tag{33}$$

Let's expand it to t-dimension:

$$\begin{aligned}&\alpha_t\alpha_{t-1}\cdots\alpha_2\beta_1\boldsymbol{\epsilon}_1+\alpha_t\alpha_{t-1}\cdots\alpha_3\beta_2\boldsymbol{\epsilon}_2\cdots+\beta_t\boldsymbol{\epsilon}_t=\bar{\boldsymbol{\epsilon}}_t\\&\bar{\boldsymbol{\epsilon}}_t\sim\mathcal{N}\left(\mathbf{0},(1-\alpha_t^2\alpha_{t-1}^2\cdots\alpha_1^2)\sigma^2\boldsymbol{I}\right),\end{aligned}\tag{34}$$

we have

$$\boldsymbol{x}_t=\alpha_t\alpha_{t-1}\cdots\alpha_1\boldsymbol{x}_0+(1-\alpha_t\alpha_{t-1}\cdots\alpha_1)\boldsymbol{\epsilon}_\mu+\sqrt{1-\alpha_t^2\alpha_{t-1}^2\cdots\alpha_1^2}\bar{\boldsymbol{\epsilon}}_t.\tag{35}$$

Since, $\bar{\alpha}_t=\alpha_t^2\alpha_{t-1}^2\cdots\alpha_1^2$, then we can get:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)=\mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu,(1-\bar{\alpha}_t)\sigma^2\boldsymbol{I}\right)\tag{36}$$

Finally, we get:

$$\begin{aligned}&\text{(Noise Process)}\qquad p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{\epsilon}_\mu)=\mathcal{N}\left(\alpha_t\boldsymbol{x}_{t-1}+(1-\alpha_t)\boldsymbol{\epsilon}_\mu,\beta_t^2\sigma^2\boldsymbol{I}\right)\\&\text{(Noise Injection)}\qquad p(\boldsymbol{x}_t|\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)=\mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu,(1-\bar{\alpha}_t)\sigma^2\boldsymbol{I}\right)\end{aligned}\tag{37}$$

$$\begin{aligned}&\text{(Noise Process)}\qquad \boldsymbol{x}_t=\alpha_t\boldsymbol{x}_{t-1}+\boldsymbol{\epsilon}_t,\boldsymbol{\epsilon}_t\sim\mathcal{N}((1-\alpha_t)\boldsymbol{\epsilon}_\mu,\beta_t^2\sigma^2\boldsymbol{I})\\&\text{(Noise Injection)}\qquad \boldsymbol{x}_t=\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+\bar{\boldsymbol{\epsilon}}_t,\bar{\boldsymbol{\epsilon}}_t\sim\mathcal{N}((1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu,(1-\bar{\alpha}_t)\sigma^2\boldsymbol{I})\end{aligned}\tag{38}$$

where $0<\alpha_t<1$ and $0<\beta_t<1$, $\sigma\in\mathbb{R}_{>0}$.

Backward Process

During the backward process, we are focus on the reverse probability distribution $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{\epsilon}_\mu)$. Hence, we have:

$$\begin{aligned}p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{\epsilon}_\mu)&=\frac{p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{\epsilon}_\mu)p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)}{p(\boldsymbol{x}_t|\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)}\\&=\frac{\mathcal{N}\left(\alpha_t\boldsymbol{x}_{t-1}+(1-\alpha_t)\boldsymbol{\epsilon}_\mu,(1-\alpha_t^2)\sigma^2\boldsymbol{I}\right)\mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0+(1-\sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\epsilon}_\mu,(1-\bar{\alpha}_{t-1})\sigma^2\boldsymbol{I}\right)}{\mathcal{N}\left(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu,(1-\bar{\alpha}_t)\sigma^2\boldsymbol{I}\right)}\end{aligned}\tag{39}$$

Due to the fact that the multi-varite gaussian distribution has probability density function $p(\boldsymbol{x})=\frac{1}{\sqrt{(2\pi)^k\det(\Sigma)}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$, where $\boldsymbol{x}\in\mathbb{R}^k$ and $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$

$$\begin{aligned}p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{\epsilon}_\mu)\propto\exp\Bigg\{-\frac{1}{2\sigma^2}\Bigg[&\frac{(\boldsymbol{x}_t-\alpha_t\boldsymbol{x}_{t-1}-(1-\alpha_t)\boldsymbol{\epsilon}_\mu)^2}{(1-\alpha_t^2)}+\frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0-(1-\sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\epsilon}_\mu)^2}{(1-\bar{\alpha}_{t-1})}\\&-\frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_t-(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu)^2}{(1-\bar{\alpha}_t)}\Bigg]\Bigg\}\end{aligned}\tag{40}$$

Since $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{\epsilon}_\mu)$ is w.r.t. $\boldsymbol{x}_{t-1}$, we can convert this pdf to a simplified formulation (by rearanging terms which has no realtionship with $\boldsymbol{x}_{t-1}$ to $C(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)$):

$$\begin{aligned}p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{\epsilon}_\mu)&\propto\exp\Bigg\{-\frac{1}{2\sigma^2}\Bigg[\frac{\boldsymbol{x}_t^2+\alpha_t^2\boldsymbol{x}_{t-1}^2+(1-\alpha)^2\boldsymbol{\epsilon}_\mu^2-2\alpha_t\boldsymbol{x}_t\boldsymbol{x}_{t-1}-2(1-\alpha_t)\boldsymbol{x}_t\boldsymbol{\epsilon}_\mu+2\alpha_t(1-\alpha_t)\boldsymbol{x}_{t-1}\boldsymbol{\epsilon}_\mu}{1-\alpha_t^2}\\&+\frac{\boldsymbol{x}_{t-1}^2+\bar{\alpha}_{t-1}\boldsymbol{x}_0^2+(1-\sqrt{\bar{\alpha}_{t-1}})^2\boldsymbol{\epsilon}_\mu^2-2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0\boldsymbol{x}_{t-1}-(2-2\sqrt{\bar{\alpha}_{t-1}})\boldsymbol{x}_{t-1}\boldsymbol{\epsilon}_\mu+2(\sqrt{\bar{\alpha}_{t-1}}+\bar{\alpha}_{t-1})\boldsymbol{x}_0\boldsymbol{\epsilon}_\mu}{1-\bar{\alpha}_{t-1}}\\&+C(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)]\}\\&=\exp\Bigg\{-\frac{1}{2\sigma^2}\Bigg[\frac{\alpha_t^2\boldsymbol{x}_{t-1}^2-2\alpha_t\boldsymbol{x}_t\boldsymbol{x}_{t-1}+2\alpha_t(1-\alpha_t)\boldsymbol{x}_{t-1}\boldsymbol{\epsilon}_\mu}{1-\alpha_t^2}+\frac{\boldsymbol{x}_{t-1}^2-2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0\boldsymbol{x}_{t-1}-(2-2\sqrt{\bar{\alpha}_{t-1}})\boldsymbol{x}_{t-1}\boldsymbol{\epsilon}_\mu}{1-\bar{\alpha}_{t-1}}\\&+C(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)]\}\end{aligned}\tag{41}$$

Then we can rerange the remaining terms to construct a gaussian distribution formulation which associated with $\boldsymbol{x}_{t-1}$.

$$
\begin{aligned}
&p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{\epsilon}_\mu)\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\frac{\alpha_t^2}{1-\alpha_t^2}+\frac{1}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}^2-2\underbrace{\left(\frac{\alpha_t\boldsymbol{x}_t-\alpha_t(1-\alpha_t)\boldsymbol{\epsilon}_\mu}{1-\alpha_t^2}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0+(1-\sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\epsilon}_\mu}{1-\bar{\alpha}_{t-1}}\right)}_{(b)}\boldsymbol{x}_{t-1}\right.\right.\\
&\qquad\qquad \left.\left.+C(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)\right]\right\}\\
&=\exp\left\{-\frac{1}{2\sigma^2}\left[\left(\frac{\alpha_t^2-\alpha_t^2\bar{\alpha}_{t-1}+1-\alpha_t^2}{(1-\alpha_t^2)(1-\bar{\alpha}_{t-1})}\right)\boldsymbol{x}_{t-1}^2-2(b)\boldsymbol{x}_{t-1}+C(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)\right]\right\}\\
&=\exp\left\{-\frac{1}{2\sigma^2}\left[\underbrace{\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t^2)(1-\bar{\alpha}_{t-1})}\right)}_{(a)}\boldsymbol{x}_{t-1}^2-2(b)\boldsymbol{x}_{t-1}+C(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)\right]\right\}\\
&=\exp\left\{-\frac{1}{2\sigma^2/(a)}\left[\boldsymbol{x}_{t-1}^2-2\frac{(b)}{(a)}\boldsymbol{x}_{t-1}+C(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)\right]\right\}\\
&=\exp\left\{-\frac{\left(\boldsymbol{x}_{t-1}-\frac{(b)}{(a)}\right)^2}{2\sigma^2(a)}+C(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)\right\}\\
&\propto\mathcal{N}\left(\boldsymbol{\mu}_p(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu),\boldsymbol{\Sigma}_p(t)\boldsymbol{I}\right),\boldsymbol{\mu}_p(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)=\frac{(b)}{(a)},\boldsymbol{\Sigma}_p(t)=\sigma^2/(a)
\end{aligned}
\tag{42}
$$

We can expand the $\boldsymbol{\mu}_p(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)$

$$
\begin{aligned}
&\boldsymbol{\mu}_p(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)\\
&=\frac{(1-\alpha_t^2)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\frac{(\alpha_t\boldsymbol{x}_t-\alpha_t(1-\alpha_t)\boldsymbol{\epsilon}_\mu)(1-\bar{\alpha}_{t-1})+(\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0+(1-\sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\epsilon}_\mu)(1-\alpha_t^2)}{(1-\alpha_t^2)(1-\bar{\alpha}_{t-1})}\\
&=\frac{(\alpha_t\boldsymbol{x}_t-\alpha_t(1-\alpha_t)\boldsymbol{\epsilon}_\mu)(1-\bar{\alpha}_{t-1})+(\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0+(1-\sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\epsilon}_\mu)(1-\alpha_t^2)}{1-\bar{\alpha}_t}\\
&=\frac{\alpha_t(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)\boldsymbol{x}_0+[\alpha_t^2-\alpha_t-\alpha_t^2\bar{\alpha}_{t-1}+\alpha_t\bar{\alpha}_{t-1}+(1-\alpha_t^2-\sqrt{\bar{\alpha}_{t-1}}+\sqrt{\bar{\alpha}_{t-1}}\alpha_t^2)]\boldsymbol{\epsilon}_\mu}{1-\bar{\alpha}_t}\\
&=\frac{\alpha_t(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)\boldsymbol{x}_0+[1-\alpha_t-\alpha_t^2\bar{\alpha}_{t-1}+\alpha_t\bar{\alpha}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}+\sqrt{\bar{\alpha}_{t-1}}\alpha_t^2]\boldsymbol{\epsilon}_\mu}{1-\bar{\alpha}_t}\\
&=\frac{\alpha_t(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)\boldsymbol{x}_0}{1-\bar{\alpha}_t}+\underbrace{\frac{1-\alpha_t-\alpha_t^2\bar{\alpha}_{t-1}+\alpha_t\bar{\alpha}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}+\sqrt{\bar{\alpha}_{t-1}}\alpha_t^2}{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\mu}_{(c)}
\end{aligned}
\tag{43}
$$

$$
\begin{aligned}
&=\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)}{1-\bar{\alpha}_t}\boldsymbol{x}_0+\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t+(c)\\
&=\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)}{1-\bar{\alpha}_t}\frac{\boldsymbol{x}_t-\mathbb{E}[\bar{\boldsymbol{\epsilon}}_t]}{\sqrt{\bar{\alpha}_t}}+\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t+(c),\bar{\boldsymbol{\epsilon}}_t\sim\mathcal{N}((1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu,(1-\bar{\alpha}_t)\sigma^2\boldsymbol{I})\\
&=\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)}{1-\bar{\alpha}_t}\frac{\boldsymbol{x}_t}{\sqrt{\bar{\alpha}_t}}+\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t-\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\mathbb{E}[\bar{\boldsymbol{\epsilon}}_t]+(c)\\
&=\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)}{1-\bar{\alpha}_t}\frac{\boldsymbol{x}_t}{\sqrt{\bar{\alpha}_t}}+\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t\underbrace{-\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)(1-\sqrt{\bar{\alpha}_t})}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_\mu}_{(d)}+(c)
\end{aligned}
$$

By reranging the terms, we can get:

$$
\begin{aligned}
\boldsymbol{\mu}_p(\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu) &= \left[\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}} + \frac{\alpha_t\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\right]\boldsymbol{x}_t + (d) + (c) \\
&= \left[\frac{1-\alpha_t^2}{(1-\bar{\alpha}_t)\alpha_t} + \frac{\alpha_t\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\right]\boldsymbol{x}_t + (d) + (c) \\
&= \frac{(1-\alpha_t^2) + \alpha_t^2(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_t)\,\alpha_t}\boldsymbol{x}_t + (d) + (c) \\
&= \frac{1-\alpha_t^2 + \alpha_t^2 - \bar{\alpha}_t}{(1-\bar{\alpha}_t)\,\alpha_t}\boldsymbol{x}_t + (d) + (c) \\
&= \frac{1}{\alpha_t}\boldsymbol{x}_t + (d) + (c)
\end{aligned} \tag{44}
$$

$$
\begin{aligned}
(c)+(d) &= \frac{(1-\alpha_t-\alpha_t^2\bar{\alpha}_{t-1}+\alpha_t\bar{\alpha}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}+\sqrt{\bar{\alpha}_{t-1}}\alpha_t^2)\sqrt{\bar{\alpha}_t}-\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)(1-\sqrt{\bar{\alpha}_t})}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_\mu \\
&= \frac{(\sqrt{\bar{\alpha}_t}-\sqrt{\bar{\alpha}_t}\alpha_t-\sqrt{\bar{\alpha}_t}\alpha_t^2\bar{\alpha}_{t-1}+\sqrt{\bar{\alpha}_t}\alpha_t\bar{\alpha}_{t-1}-\sqrt{\bar{\alpha}_t}\sqrt{\bar{\alpha}_{t-1}}+\sqrt{\bar{\alpha}_t}\sqrt{\bar{\alpha}_{t-1}}\alpha_t^2)-\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)(1-\sqrt{\bar{\alpha}_t})}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_\mu \\
&= \frac{(\sqrt{\bar{\alpha}_t}-\sqrt{\bar{\alpha}_t}\alpha_t-\sqrt{\bar{\alpha}_t}\alpha_t^2\bar{\alpha}_{t-1}+\sqrt{\bar{\alpha}_t}\alpha_t\bar{\alpha}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}+\sqrt{\bar{\alpha}_{t-1}}\alpha_t^2)}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_\mu \\
&= \frac{(\sqrt{\bar{\alpha}_t}-\sqrt{\bar{\alpha}_t}\alpha_t^2\bar{\alpha}_{t-1}+\sqrt{\bar{\alpha}_t}\alpha_t\bar{\alpha}_{t-1}-\sqrt{\bar{\alpha}_{t-1}})}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_\mu \\
&= \frac{(\alpha_t\sqrt{\bar{\alpha}_{t-1}}-\sqrt{\bar{\alpha}_t}\alpha_t^2\bar{\alpha}_{t-1}+\sqrt{\bar{\alpha}_t}\alpha_t\bar{\alpha}_{t-1}-\sqrt{\bar{\alpha}_{t-1}})}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_\mu \\
&= \frac{(1-\alpha_t)(\sqrt{\bar{\alpha}_t}\alpha_t\bar{\alpha}_{t-1}-\sqrt{\bar{\alpha}_{t-1}})}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_\mu \\
&= \frac{(1-\alpha_t)(\sqrt{\bar{\alpha}_t}\bar{\alpha}_t/\alpha_t-\sqrt{\bar{\alpha}_t}/\alpha_t)}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{\epsilon}_\mu \\
&= \frac{(1-\alpha_t)(\bar{\alpha}_t-1)}{(1-\bar{\alpha}_t)\alpha_t}\boldsymbol{\epsilon}_\mu \\
&= \frac{\alpha_t-1}{\alpha_t}\boldsymbol{\epsilon}_\mu
\end{aligned} \tag{45}
$$

Then we can get:

$$
\boldsymbol{\mu}_p(\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu) = \frac{1}{\alpha_t}\boldsymbol{x}_t - \frac{1-\alpha_t}{\alpha_t}\boldsymbol{\epsilon}_\mu \tag{46}
$$

Finally, we have

$$
\text{(Noise Process)} \qquad p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{\epsilon}_\mu) = \mathcal{N}\left(\frac{1}{\alpha_t}\boldsymbol{x}_t - \frac{1-\alpha_t}{\alpha_t}\boldsymbol{\epsilon}_\mu, \left(\frac{(1-\alpha_t^2)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\right)\sigma^2\boldsymbol{I}\right) \tag{47}
$$

$$
\text{(Noise Process)} \qquad \boldsymbol{x}_{t-1} = \frac{1}{\alpha_t}\boldsymbol{x}_t - \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \sim \mathcal{N}\left(\frac{1-\alpha_t}{\alpha_t}\boldsymbol{\epsilon}_\mu, \left(\frac{(1-\alpha_t^2)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\right)\sigma^2\boldsymbol{I}\right) \tag{48}
$$

□

A.2. Probability Path

Proposition 2. *For any general diffusion which defined in Prop.1, we have the path probability (the probability of starting from $\hat{\boldsymbol{x}}_0$ forward to $\hat{\boldsymbol{\epsilon}}_\mu$ and backward to $\hat{\boldsymbol{x}}_0$)*

$$
p(\hat{\boldsymbol{x}}_0 \to \hat{\boldsymbol{\epsilon}}_\mu \to \hat{\boldsymbol{x}}_0) = p(\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu|\boldsymbol{x}_0 = \hat{\boldsymbol{x}}_0)p(\boldsymbol{x}_0 = \hat{\boldsymbol{x}}_0|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu) \tag{49}
$$

where

$$p(\boldsymbol{x}_0|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)=\mathcal{N}\left(\frac{\hat{\boldsymbol{\epsilon}}_\mu}{\sqrt{\bar{\alpha}_T}}+\sum_{s=1}^{T}\frac{(e)_s+(c)_s}{\sqrt{\bar{\alpha}_{s-1}/\bar{\alpha}_{t-1}}},\sum_{s=1}^{T}\frac{\tilde{\beta}_s\bar{\alpha}_t}{\bar{\alpha}_s}\sigma^2\boldsymbol{I}\right) \tag{50}$$

and

$$(e)_s=-\frac{\sqrt{\bar{\alpha}_{s-1}}(1-\alpha_s^2)}{(1-\bar{\alpha}_s)\sqrt{\bar{\alpha}_s}}\mathbb{E}[\bar{\boldsymbol{\epsilon}}_s] \tag{51}$$

$$(c)_s=\frac{1-\alpha_s-\alpha_s^2\bar{\alpha}_{s-1}+\alpha_t\bar{\alpha}_{s-1}-\sqrt{\bar{\alpha}_{s-1}}+\sqrt{\bar{\alpha}_{s-1}}\alpha_s^2}{1-\bar{\alpha}_s}\boldsymbol{\epsilon}_\mu \tag{52}$$

Proof. According to Eq.(43), we have

$$\begin{aligned}
\boldsymbol{\mu}_p(\boldsymbol{x}_t,\boldsymbol{x}_0,\boldsymbol{\epsilon}_\mu)&=\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)}{1-\bar{\alpha}_t}\frac{\boldsymbol{x}_t-\mathbb{E}[\bar{\boldsymbol{\epsilon}}_t]}{\sqrt{\bar{\alpha}_t}}+\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t+(c)_t,\bar{\boldsymbol{\epsilon}}_t\sim\mathcal{N}((1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu,(1-\bar{\alpha}_t)\sigma^2\boldsymbol{I})\\
&=\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\boldsymbol{x}_t+\frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t\underbrace{-\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t^2)\mathbb{E}[\bar{\boldsymbol{\epsilon}}_t]}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}}_{(e)_t}+(c)_t\\
&=\frac{1}{\alpha_t}\boldsymbol{x}_t+(e)_t+(c)_t,
\end{aligned} \tag{53}$$

Then we can write the backward process probability:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)=\mathcal{N}\left(\frac{1}{\alpha_t}\boldsymbol{x}_t+(e)_t+(c)_t,\tilde{\beta}_t\sigma^2\boldsymbol{I}\right) \tag{54}$$

For the convenience of subsequent analysis, we start from timestep T

$$p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)=\mathcal{N}\left(\frac{1}{\alpha_T}\hat{\boldsymbol{\epsilon}}_\mu+(e)_T+(c)_T,\tilde{\beta}_T\sigma^2\boldsymbol{I}\right) \tag{55}$$

Then we can obtain the marginal distribution $p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)$.

$$\begin{aligned}
p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)&=\int_{\mathbb{R}^n}p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1})p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)d\boldsymbol{x}_{T-1}\\
&=\int_{\mathbb{R}^n}\mathcal{N}\left(\frac{\boldsymbol{x}_{T-1}}{\alpha_{T-1}}+(e)_{T-1}+(c)_{T-1},\tilde{\beta}_{T-1}\sigma^2\boldsymbol{I}\right)\mathcal{N}\left(\frac{1}{\alpha_T}\hat{\boldsymbol{\epsilon}}_\mu+(e)_T+(c)_T,\tilde{\beta}_T\sigma^2\boldsymbol{I}\right)d\boldsymbol{x}_{T-1}
\end{aligned} \tag{56}$$

It's hard to calculate $p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)$ directly. However, we can ensure that $p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)$ is a gaussian distribution. Hence, we just need to calculate the paramter of gaussian distribution (mean and variance).

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu]\\
&=\mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}[\mathbb{E}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1}]|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu]\\
&=\mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}\left[\mathbb{E}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}\left[\frac{\boldsymbol{x}_{T-1}}{\alpha_{T-1}}+(e)_{T-1}+(c)_{T-1}|\boldsymbol{x}_{T-1}\right]|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu\right]\\
&=\mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}\left[\frac{1}{\alpha_{T-1}}\left(\frac{1}{\alpha_T}\hat{\boldsymbol{\epsilon}}_\mu+(e)_T+(c)_T\right)+(e)_{T-1}+(c)_{T-1}\right]\\
&=\frac{\hat{\boldsymbol{\epsilon}}_\mu}{\alpha_T\alpha_{T-1}}+\frac{(e)_T+(c)_T}{\alpha_{T-1}}+(e)_{T-1}+(c)_{T-1}
\end{aligned} \tag{57}$$

Due to the fact that $\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$, hence, we can caculate the variance by:

$$\begin{aligned}
&\mathbb{V}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu] \\
&= \mathbb{E}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}\left[\mathbb{V}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1})}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1}]|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu\right] \\
&\quad + \mathbb{V}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}\left[\mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1})}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1}]|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu\right] \\
&= \mathbb{E}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}\left[\tilde{\beta}_{T-1}\sigma^2\boldsymbol{I}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu\right] \\
&\quad + \mathbb{V}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)}\left[\frac{1}{\alpha_{T-1}}\boldsymbol{x}_{T-1} - \frac{1-\alpha_{T-1}}{\alpha_{T-1}}\boldsymbol{\epsilon}_\mu|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu\right] \\
&= \tilde{\beta}_{T-1}\sigma^2\boldsymbol{I} + \left(\frac{1}{\alpha_{T-1}}\right)^2\tilde{\beta}_T\sigma^2\boldsymbol{I} \\
&= \left(\tilde{\beta}_{T-1} + \frac{\tilde{\beta}_T}{\alpha_{T-1}^2}\right)\sigma^2\boldsymbol{I}
\end{aligned} \tag{58}$$

where $\tilde{\beta}_t = \left(\frac{(1-\alpha_t^2)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\right)$, After obtaining the paramter of $p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)$, we can get

$$p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu) = \mathcal{N}\left(\frac{\hat{\boldsymbol{\epsilon}}_\mu}{\sqrt{\bar{\alpha}_T/\bar{\alpha}_{T-2}}} + \frac{(e)_T+(c)_T}{\alpha_{T-1}} + (e)_{T-1} + (c)_{T-1}, \left(\tilde{\beta}_{T-1} + \frac{\tilde{\beta}_T}{\alpha_{T-1}^2}\right)\sigma^2\boldsymbol{I}\right) \tag{59}$$

It's easy to get the general term $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)$ by mathematical induction.

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu) = \mathcal{N}\left(\frac{\hat{\boldsymbol{\epsilon}}_\mu}{\sqrt{\bar{\alpha}_T/\bar{\alpha}_{t-1}}} + \sum_{s=t}^{T}\frac{(e)_s+(c)_s}{\sqrt{\bar{\alpha}_{s-1}/\bar{\alpha}_{t-1}}}, \sum_{s=t}^{T}\frac{\tilde{\beta}_s\bar{\alpha}_t}{\bar{\alpha}_s}\sigma^2\boldsymbol{I}\right) \tag{60}$$

Finally, we can inference the probability density distribution at the timestep 0.

$$p(\boldsymbol{x}_0|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu) = \mathcal{N}\left(\frac{\hat{\boldsymbol{\epsilon}}_\mu}{\sqrt{\bar{\alpha}_T}} + \sum_{s=1}^{T}\frac{(e)_s+(c)_s}{\sqrt{\bar{\alpha}_{s-1}/\bar{\alpha}_{t-1}}}, \sum_{s=1}^{T}\frac{\tilde{\beta}_s\alpha_1}{\bar{\alpha}_s}\sigma^2\boldsymbol{I}\right) \tag{61}$$

□

A.3. Optimal Probability Path

Theorem 1. *Let $\boldsymbol{S}_{\bar{\epsilon}} = \{\bar{\boldsymbol{\epsilon}}_t\}_{t=1}^T$ be the noise added in the forward process, $\boldsymbol{\epsilon}_\mu$, σ be the parameters of the target distribution $\mathcal{N}(\boldsymbol{\epsilon}_\mu, \sigma^2\boldsymbol{I})$. For sample $\hat{\boldsymbol{x}}_0$, we can obtain the optimal parameters that maximize the path probability.*

$$\underset{\boldsymbol{S}_{\bar{\epsilon}}, \hat{\boldsymbol{\epsilon}}_\mu, \boldsymbol{\epsilon}_\mu, \sigma}{\arg\max}\; p(\boldsymbol{x}_0=\hat{\boldsymbol{x}}_0|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu)\prod_{t=1}^{T}p(\bar{\boldsymbol{\epsilon}}_t=\mathbb{E}[\bar{\boldsymbol{\epsilon}}_t]) \tag{62}$$

where the optimal parameters are defined by:

$$\begin{aligned}
&\boldsymbol{S}_{\bar{\epsilon}}^* = \{(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu\}_{t=1}^T,\ \sigma^* \to 0 \\
&\hat{\boldsymbol{\epsilon}}_\mu^* = \sqrt{\bar{\alpha}_T}\hat{\boldsymbol{x}}_0 + (1-\sqrt{\bar{\alpha}_T})\boldsymbol{\epsilon}_\mu,\ \boldsymbol{\epsilon}_\mu^* = \mathbb{E}_t[\mathbb{E}[\bar{\boldsymbol{\epsilon}}_t]]
\end{aligned} \tag{63}$$

Proof. According to Eq.(45), we can rewrite the backward process probability distribution.

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*) = \mathcal{N}\left(\frac{1}{\alpha_t}\boldsymbol{x}_t - \frac{1-\alpha_T}{\alpha_T}\boldsymbol{\epsilon}_\mu^*, \tilde{\beta}_t(\sigma^*)^2\boldsymbol{I}\right) \tag{64}$$

For the convenience of subsequent analysis, we start from timestep T

$$p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*) = \mathcal{N}\left(\frac{1}{\alpha_T}\hat{\boldsymbol{\epsilon}}_\mu^* - \frac{1-\alpha_T}{\alpha_T}\boldsymbol{\epsilon}_\mu^*, \tilde{\beta}_T(\sigma^*)^2\boldsymbol{I}\right) \tag{65}$$

Then we can obtain the marginal distribution $p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*)$.

$$\begin{aligned} p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*) &= \int_{\mathbb{R}^n} p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1}) p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*) d\boldsymbol{x}_{T-1} \\ &= \int_{\mathbb{R}^n} \mathcal{N}\left(\frac{\boldsymbol{x}_{T-1}}{\alpha_{T-1}} - \frac{1-\alpha_{T-1}}{\alpha_{T-1}}\boldsymbol{\epsilon}_\mu^*, \tilde{\beta}_{T-1}(\sigma^*)a^2\boldsymbol{I}\right)\mathcal{N}\left(\frac{1}{\alpha_T}\hat{\boldsymbol{\epsilon}}_\mu - \frac{1-\alpha_T}{\alpha_T}\boldsymbol{\epsilon}_\mu^*, \tilde{\beta}_T(\sigma^*)^2\boldsymbol{I}\right) d\boldsymbol{x}_{T-1} \end{aligned} \tag{66}$$

Similar to Prop.2. It's hard to calculate $p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*)$ directly. However, we can ensure that $p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*)$ is a gaussian distribution. Hence, we just need to calculate the paramter of gaussian distribution (mean and variance).

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*] \\ =& \mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}[\mathbb{E}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1}]|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*] \\ =& \mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}\left[\mathbb{E}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}\left[\frac{\boldsymbol{x}_{T-1}}{\alpha_{T-1}} - \frac{1-\alpha_{T-1}}{\alpha_{T-1}}\boldsymbol{\epsilon}_\mu^*|\boldsymbol{x}_{T-1}\right]|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*\right] \\ =& \mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}\left[\frac{1}{\alpha_{T-1}}\left(\frac{1}{\alpha_T}\hat{\boldsymbol{\epsilon}}_\mu^* - \frac{1-\alpha_T}{\alpha_T}\boldsymbol{\epsilon}_\mu^*\right) - \frac{1-\alpha_{T-1}}{\alpha_{T-1}}\boldsymbol{\epsilon}_\mu^*|\boldsymbol{x}_{T-1}\right] \\ =& \frac{\hat{\boldsymbol{\epsilon}}_\mu^*}{\alpha_T\alpha_{T-1}} - \frac{1-\alpha_T\alpha_{T-1}}{\alpha_T\alpha_{T-1}}\boldsymbol{\epsilon}_\mu^* \\ =& \frac{\hat{\boldsymbol{\epsilon}}_\mu^*}{\sqrt{\bar{\alpha}_T/\bar{\alpha}_{T-2}}} - \frac{1-\sqrt{\bar{\alpha}_T/\bar{\alpha}_{T-2}}}{\sqrt{\bar{\alpha}_T/\bar{\alpha}_{T-2}}}\boldsymbol{\epsilon}_\mu^* \end{aligned} \tag{67}$$

and for the variance, we have

$$\begin{aligned} &\mathbb{V}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*] \\ =& \mathbb{E}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}\left[\mathbb{V}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1})}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1}]|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*\right] \\ &+ \mathbb{V}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}\left[\mathbb{E}_{\boldsymbol{x}_{T-2}\sim p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1})}[\boldsymbol{x}_{T-2}|\boldsymbol{x}_{T-1}]|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*\right] \\ =& \mathbb{E}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}\left[\tilde{\beta}_{T-1}\sigma^2\boldsymbol{I}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*\right] \\ &+ \mathbb{V}_{\boldsymbol{x}_{T-1}\sim p(\boldsymbol{x}_{T-1}|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu^*)}\left[\frac{1}{\alpha_{T-1}}\boldsymbol{x}_{T-1} - \frac{1-\alpha_{T-1}}{\alpha_{T-1}}\boldsymbol{\epsilon}_\mu^*|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*\right] \\ =& \tilde{\beta}_{T-1}(\sigma^*)^2\boldsymbol{I} + \left(\frac{1}{\alpha_{T-1}}\right)^2\tilde{\beta}_T(\sigma^*)^2\boldsymbol{I} \\ =& \left(\tilde{\beta}_{T-1} + \frac{\tilde{\beta}_T}{\alpha_{T-1}^2}\right)(\sigma^*)^2\boldsymbol{I} \end{aligned} \tag{68}$$

After obtaining the paramter of $p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu^*)$, we can get

$$p(\boldsymbol{x}_{T-2}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu) = \mathcal{N}\left(\frac{\hat{\boldsymbol{\epsilon}}_\mu}{\sqrt{\bar{\alpha}_T/\bar{\alpha}_{T-2}}} - \frac{1-\sqrt{\bar{\alpha}_T/\bar{\alpha}_{T-2}}}{\sqrt{\bar{\alpha}_T/\bar{\alpha}_{T-2}}}\boldsymbol{\epsilon}_\mu, \left(\tilde{\beta}_{T-1} + \frac{\tilde{\beta}_T}{\alpha_{T-1}^2}\right)\sigma^2\boldsymbol{I}\right) \tag{69}$$

It's easy to get the general term $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu)$ by mathematical induction.

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_T = \hat{\boldsymbol{\epsilon}}_\mu) = \mathcal{N}\left(\frac{\hat{\boldsymbol{\epsilon}}_\mu}{\sqrt{\bar{\alpha}_T/\bar{\alpha}_{t-1}}} - \frac{1-\sqrt{\bar{\alpha}_T/\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_T/\bar{\alpha}_{t-1}}}\boldsymbol{\epsilon}_\mu, \sum_{s=t}^{T}\frac{\tilde{\beta}_s\bar{\alpha}_t}{\bar{\alpha}_s}\sigma^2\boldsymbol{I}\right) \tag{70}$$

Finally, we can inference the probability density distribution at the timestep 0.

$$\begin{aligned} p(\boldsymbol{x}_0|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}^*_\mu) &= \mathcal{N}\left(\frac{\hat{\boldsymbol{\epsilon}}^*_\mu}{\sqrt{\bar{\alpha}_T}} - \frac{1-\sqrt{\bar{\alpha}_T}}{\sqrt{\bar{\alpha}_T}}\boldsymbol{\epsilon}^*_\mu, \sum_{s=t}^{T}\frac{\tilde{\beta}_s\bar{\alpha}_t}{\bar{\alpha}_s}(\sigma^*)^2\boldsymbol{I}\right) \\ &= \mathcal{N}\left(\frac{\sqrt{\bar{\alpha}_T}\hat{\boldsymbol{x}}_0+(1-\sqrt{\bar{\alpha}_T})\boldsymbol{\epsilon}^*_\mu}{\sqrt{\bar{\alpha}_T}} - \frac{1-\sqrt{\bar{\alpha}_T}}{\sqrt{\bar{\alpha}_T}}\boldsymbol{\epsilon}^*_\mu, \sum_{s=t}^{T}\frac{\tilde{\beta}_s\bar{\alpha}_t}{\bar{\alpha}_s}(\sigma^*)^2\boldsymbol{I}\right) \\ &= \mathcal{N}\left(\hat{\boldsymbol{x}}_0, \sum_{s=t}^{T}\frac{\tilde{\beta}_s\bar{\alpha}_t}{\bar{\alpha}_s}(\sigma^*)^2\boldsymbol{I}\right) = \mathcal{N}(\boldsymbol{0},\boldsymbol{I}) \end{aligned} \tag{71}$$

Hence, we can induce

$$p(\boldsymbol{x}_0=\hat{\boldsymbol{x}}_0|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}^*_\mu) = 1 \geq p(\boldsymbol{x}_0=\hat{\boldsymbol{x}}_0|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}_\mu) \tag{72}$$

$p(\boldsymbol{x}_0=\hat{\boldsymbol{x}}_0|\boldsymbol{x}_T=\hat{\boldsymbol{\epsilon}}^*_\mu)$ is the optimal probability path, and $\boldsymbol{S}^*_{\boldsymbol{\epsilon}}, \hat{\boldsymbol{\epsilon}}^*_\mu, \boldsymbol{\epsilon}^*_\mu, \sigma^*$ are optimal parameters. □

A.4. Optimal Denoiser

Let us assume that out training set consists of a finite number of samples $\{\boldsymbol{x}_0^{(1)},\cdots,\boldsymbol{x}_0^{(K)}\}$ and target mean set $\{\boldsymbol{\epsilon}_\mu^{(1)},\cdots,\boldsymbol{\epsilon}_\mu^{(K)}\}$. This implies $p(\boldsymbol{x}_0)$ is representsed by a mixture of Dirac delta distributions:

$$p(\boldsymbol{x}_0) = \frac{1}{K}\sum_{i=1}^{K}\delta(\boldsymbol{x}_0 - \boldsymbol{x}_0^{(i)}) \tag{73}$$

Let us now consider the denoising loss. By expanding the expectations, we can rewrite the formula as an integral over the noisy samples $\boldsymbol{x}$

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}_0^{(i)}\sim p(\boldsymbol{x}_0)}[\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t)-\boldsymbol{\epsilon}_\mu^{(i)}\|_2^2]] \\ &= \mathbb{E}_{\boldsymbol{x}_0^{(i)}\sim p(\boldsymbol{x}_0)}\left[\int_{\mathbb{R}^n}\mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)}+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)},(1-\bar{\alpha}_t)\boldsymbol{I})\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t)-\boldsymbol{\epsilon}_\mu^{(i)}\|_2^2\,d\boldsymbol{x}_t\right] \\ &= \frac{1}{K}\sum_{i=1}^{K}\int_{\mathbb{R}^n}\mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)}+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)},(1-\bar{\alpha}_t)\boldsymbol{I})\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t)-\boldsymbol{\epsilon}_\mu^{(i)}\|_2^2\,d\boldsymbol{x}_t \\ &= \int_{\mathbb{R}^n}\underbrace{\frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)}+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)},(1-\bar{\alpha}_t)\boldsymbol{I})\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t)-\boldsymbol{\epsilon}_\mu^{(i)}\|_2^2}_{\mathcal{L}(\boldsymbol{\epsilon}_\theta)}\,d\boldsymbol{x}_t \end{aligned} \tag{74}$$

We can minimize $\mathcal{L}(\boldsymbol{\epsilon}_\theta)$ independently for each $\boldsymbol{x}_t$:

$$\boldsymbol{\epsilon}^*_\theta = \arg\min_{\boldsymbol{\epsilon}_\theta}\mathcal{L}(\boldsymbol{\epsilon}_\theta) \tag{75}$$

This is a convex optimization problem; its solution is uniquely identified by setting the gradient w.r.t. $\boldsymbol{\epsilon}_\theta$ to zero:

$$\begin{aligned} &\nabla_{\boldsymbol{\epsilon}_\theta}\mathcal{L}(\boldsymbol{\epsilon}_\theta) \\ &= \nabla_{\boldsymbol{\epsilon}_\theta}\left[\frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)}+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)},(1-\bar{\alpha}_t)\boldsymbol{I})\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t)-\boldsymbol{\epsilon}_\mu^{(i)}\|_2^2\right] \\ &= \left[\frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)}+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)},(1-\bar{\alpha}_t)\boldsymbol{I})\right]\nabla_{\boldsymbol{\epsilon}_\theta}\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t)-\boldsymbol{\epsilon}_\mu^{(i)}\|_2^2 \\ &= \left[\frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)}+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)},(1-\bar{\alpha}_t)\boldsymbol{I})\right](2\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t)-2\boldsymbol{\epsilon}_\mu^{(i)}) \\ &= \left[\frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)}+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)},(1-\bar{\alpha}_t)\boldsymbol{I})\right]\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t) - \left[\frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)}+(1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)},(1-\bar{\alpha}_t)\boldsymbol{I})\right]\boldsymbol{\epsilon}_\mu^{(i)} \end{aligned} \tag{76}$$

Then we can solve the optimal denoiser by letting $\nabla_{\boldsymbol{\epsilon}_\theta}\mathcal{L}(\boldsymbol{\epsilon}_\theta) = 0$

$$\boldsymbol{\epsilon}_\theta^*(\boldsymbol{x}_t) = \frac{\frac{1}{K}\sum_{i=1}^K \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)} + (1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)}, (1-\bar{\alpha}_t)\boldsymbol{I})\boldsymbol{\epsilon}_\mu^{(i)}}{\frac{1}{K}\sum_{i=1}^K \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{(i)} + (1-\sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu^{(i)}, (1-\bar{\alpha}_t)\boldsymbol{I})} \tag{77}$$

B. Timestep Importance Sampling

B.1. Stein Identity

The KSD is thus given by the norm of $\boldsymbol{\beta}$ in $\mathcal{H}^d$:

Lemma 1. *For any bounded and smooth function $f(x)$, we have $\mathbb{E}_p[s_q(x)f(x) + \nabla_x f(x)] = 0$, where $s_q(x) = \nabla_x \ln q(x)$ is the score function of distribution q.*

Proof. Expanding the expectation:

$$\int s_q(x)f(x)p(x)dx + \int \nabla_x f(x)p(x)dx = 0 \tag{78}$$

Let's focus on the first term using integration by parts:

$$\int s_q(x)f(x)p(x)dx = \int \nabla_x \ln q(x)f(x)p(x)dx \tag{79}$$

Let $u = f(x)$ and $dv = \nabla_x \ln q(x)p(x)dx$. Then $du = \nabla_x f(x)dx$ and $v = \int \nabla_x \ln q(x)p(x)dx$.
Applying integration by parts:

$$\int s_q(x)f(x)p(x)dx = \left[f(x)\int \nabla_x \ln q(x)p(x)dx\right]_{-\infty}^{\infty} - \int \nabla_x f(x)\left(\int \nabla_x \ln q(x)p(x)dx\right)dx \tag{80}$$

Let $A = \int \nabla_x \ln q(x)p(x)dx$. Substituting back into Eq.(78):

$$\int s_q(x)f(x)p(x)dx + \int \nabla_x f(x)p(x)dx = [f(x)A]_{-\infty}^{\infty} - \int \nabla_x f(x)(p(x) - A)dx \tag{81}$$

Now, let's examine the two terms on the right-hand side:
1. $[f(x)A]_{-\infty}^{\infty}$: This term vanishes because $f(x)$ is bounded and $q(x)$ approaches zero at infinity, implying A also approaches zero at infinity.
2. $\int \nabla_x f(x)(p(x) - A)dx$: This term equals zero if and only if $p(x) = A$ for all x.
Since A is a constant, we have:

$$\nabla_x p(x) = \nabla_x A = 0 \tag{82}$$

This implies:

$$\frac{\nabla_x p(x)}{p(x)} = \nabla_x \ln q(x) \tag{83}$$

Which is true if and only if $p(x) = q(x)$, as $\nabla_x \ln p(x) = \frac{\nabla_x p(x)}{p(x)}$.
Therefore, when $p(x) = q(x)$, both terms on the right-hand side are zero, proving the theorem. □

B.2. Stein Discrepancy

Let's develop the concept of Stein Discrepancy as a measure of distance between probability distributions.

B.2.1 Basic Definition

Consider two smooth distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ on $\mathbb{R}^d$. The Stein score function for distribution q is defined as:

$$s_q(\mathbf{x}) = \nabla_{\mathbf{x}} \ln q(\mathbf{x}) \tag{84}$$

A fundamental property (proved in Sec.B.1) states that:

$$\mathbb{E}p[s_q(\mathbf{x})f(\mathbf{x}) + \nabla\mathbf{x}f(\mathbf{x})] = 0 \tag{85}$$

holds if and only if $p(\mathbf{x}) = q(\mathbf{x})$, where f is a smooth function.

B.2.2 Stein Discrepancy Measure

Based on this property, we can define a discrepancy measure between distributions:

$$S(p, q) = \max_{f \in \mathcal{F}} (\mathbb{E}p[s_q(\mathbf{x})f(\mathbf{x}) + \nabla \mathbf{x} f(\mathbf{x})])^2 \tag{86}$$

where $\mathcal{F}$ is the set of smooth functions. This measure is positive when $p \neq q$, but is generally difficult to compute directly.

B.2.3 RKHS Framework

To make the discrepancy measure computationally tractable, we can utilize Reproducing Kernel Hilbert Space (RKHS) theory. Let's recall key RKHS concepts:

For a positive definite kernel $k(\mathbf{x}, \mathbf{y})$, Mercer's theorem gives its spectral decomposition:

$$k(\mathbf{x}, \mathbf{y}) = \sum_j \lambda_j \mathbf{e}_j(\mathbf{x}) \mathbf{e}_j(\mathbf{y})^T \tag{87}$$

where $\mathbf{e}_j$ are orthogonal basis functions and λ_j are eigenvalues. The RKHS $\mathcal{H}$ generated by kernel $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has two key properties: $k(\mathbf{x}, \cdot) \in \mathcal{H}$ for any $\mathbf{x} \in \mathcal{X}$ Reproducing property: $f(\mathbf{x}) = \langle f, k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$ We extend this to $\mathcal{H}^d = \mathcal{H} \times \cdots \times \mathcal{H}$ (d times) for vector-valued functions with inner product:

$$\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{H}^d} = \sum_{i=1}^{d} \langle f_i, g_i \rangle_{\mathcal{H}} \tag{88}$$

B.2.4 Kernelized Stein Discrepancy (KSD)

The Stein operator $\mathcal{A}p$ for distribution p is defined as:

$$\mathcal{A}p f(\mathbf{x}) = s_p(\mathbf{x}) \cdot f(\mathbf{x}) + \nabla \mathbf{x} \cdot f(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \nabla \mathbf{x} \cdot [p(\mathbf{x}) f(\mathbf{x})] \tag{89}$$

This operator satisfies:

$$\int_{\mathbf{x} \in \mathcal{X}} \nabla_{\mathbf{x}} \cdot (f(\mathbf{x}) p(\mathbf{x})) d\mathbf{x} = 0 \tag{90}$$

A key relationship for KSD is:

$$\mathbb{E}_p[\mathcal{A}_q \mathbf{f}(\mathbf{x})] = \mathbb{E}_p[\mathcal{A}_q \mathbf{f}(\mathbf{x}) - \mathcal{A}_p \mathbf{f}(\mathbf{x})] = \mathbb{E}_p[(s_q(\mathbf{x}) - s_p(\mathbf{x})) \mathbf{f}(\mathbf{x})^T] \tag{91}$$

This formulation provides a computationally tractable way to measure discrepancy between distributions using kernel methods.

B.3. Optimal sampling Distribution

Proposition 3. *The optimal sampling distribution for Eq.(5) with minimal variance is:*

$$q^*(t|\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu) \propto \xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu) p(t), \tag{92}$$

where $\xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu) = \|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\hat{\boldsymbol{x}}_0 + (1 - \sqrt{\bar{\alpha}_t})\boldsymbol{\epsilon}_\mu) - \boldsymbol{\epsilon}_\mu\|_2^2$. *which means for any probability distribution* p*, we have*

$$\mathbb{V}_{t \sim q^*(t), (\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)}[\xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)] \leq \mathbb{V}_{t \sim p(t), (\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)}[\xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)]$$

Proof. We aim to minimize the variance:

$$\mathbb{V}_{t \sim q(t), (\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)} \left[\xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)\right]$$

This can be expressed as:

$$\int \frac{\xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu) p(t)^2}{q(t)} \, dt$$

subject to $\int q(t)\,dt = 1$. The Lagrangian is:

$$\int \frac{\xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)p(t)^2}{q(t)}\,dt + \lambda\left(\int q(t)\,dt - 1\right)$$

Taking the derivative with respect to $q(t)$ and setting it to zero gives:

$$-\frac{\xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)p(t)^2}{q(t)^2} + \lambda = 0$$

Solving for $q(t)$, we get:

$$q^*(t) = \sqrt{\frac{\xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)p(t)^2}{\lambda}}$$

This implies:

$$q^*(t) \propto \xi_t(\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)p(t)$$

Using the Cauchy-Schwarz inequality, we show:

$$\sigma_{q^*}^2 \leq \sigma_p^2$$

Thus, the distribution $q^*(t|\boldsymbol{x}_0, \boldsymbol{\epsilon}_\mu)$ minimizes the variance for the given problem, proving the proposition. □