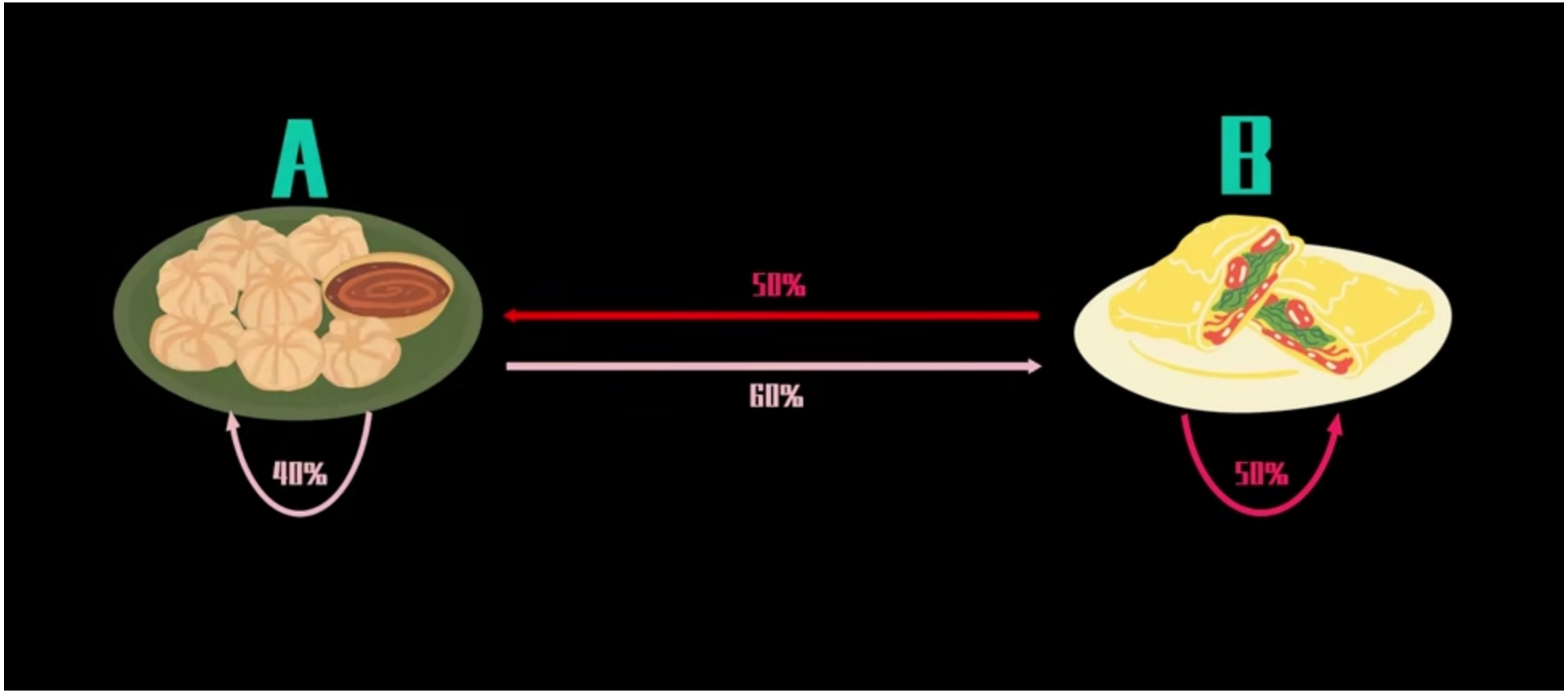
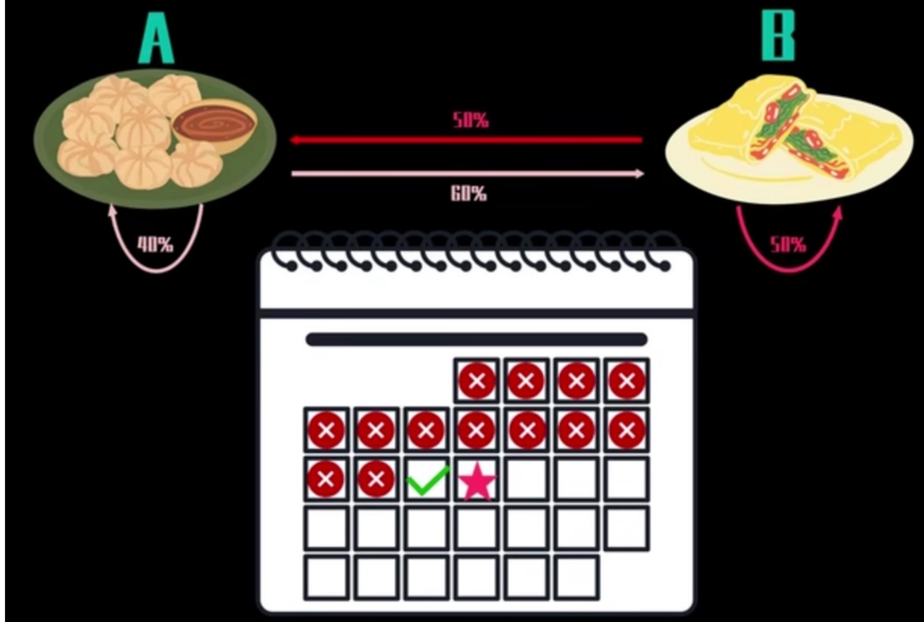
DDIM


A
B
50%
60%
40%
50%



马尔可夫链的核心三要素:

1. 状态空间 States Space

2. 无记忆性 Memorylessness

   $P(S_t \mid S_{t-1}, S_{t-2}, S_{t-3}, \ldots\ldots) = P(S_t \mid S_{t-1})$

3. 转移矩阵 Transition Matrix

	A（前日）	B（前日）
A（当日）	0.4	0.5
B（当日）	0.6	0.5

A B

$$\begin{bmatrix}0.4 & 0.5\\ 0.6 & 0.5\end{bmatrix} \times \begin{bmatrix}1\\ 0\end{bmatrix}\begin{array}{c}A\\ B\end{array} = \begin{bmatrix}0.4\\ 0.6\end{bmatrix}\begin{array}{c}A\\ B\end{array}$$

$$\begin{bmatrix}0.4 & 0.5\\ 0.6 & 0.5\end{bmatrix}^{2} \times \begin{bmatrix}1\\ 0\end{bmatrix}\begin{array}{c}A\\ B\end{array} = \begin{bmatrix}0.46\\ 0.54\end{bmatrix}$$

$$\begin{bmatrix}0.4 & 0.5\\ 0.6 & 0.5\end{bmatrix}^{n} \times \begin{bmatrix}1\\ 0\end{bmatrix}\begin{array}{c}A\\ B\end{array} = \begin{bmatrix}0.454545\\ 0.545455\end{bmatrix}$$

$$\begin{bmatrix}0.4 & 0.5\\ 0.6 & 0.5\end{bmatrix}^{n} \times \begin{bmatrix}0\\ 1\end{bmatrix}\begin{array}{c}A\\ B\end{array} = \begin{bmatrix}0.454545\\ 0.545455\end{bmatrix}$$

回顾DDPM

前向过程 ☆ $$\boxed{q(x_t|x_0)=N(x_t;\sqrt{\bar\alpha_t}x_0,(1-\bar\alpha_t)I)}$$

$$x_t=\sqrt{\bar\alpha_t}x_0+\sqrt{1-\bar\alpha_t}\,\varepsilon \qquad \varepsilon\sim N(0,I)$$

$$q(x_{t-1}|x_t)=\frac{q(x_t|x_{t-1})\,q(x_{t-1})}{q(x_t)}$$

$$\parallel$$

逆向过程 ☆ $$\boxed{q(x_{t-1}|x_t,x_0)=\frac{q(x_t|x_{t-1})\,q(x_{t-1}|x_0)}{q(x_t|x_0)}=N(x_{t-1};\mu(x_t,x_0),\sigma_t^2 I)}$$

$$q(x_t|x_{t-1})=N(x_t;\sqrt{\alpha_t}x_{t-1},(1-\alpha_t)I)$$

$$\hat{x}_{0|t}=\frac{1}{\sqrt{\bar\alpha_t}}(x_t-\sqrt{1-\bar\alpha_t}\,\varepsilon_\theta)$$

$$\mu(x_t,x_0)=\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}x_t+\frac{\sqrt{\bar\alpha_{t-1}}\,\beta_t}{1-\bar\alpha_t}x_0$$

$$\sigma_t^2=\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\cdot\beta_t$$

一窥DDIM

$$q(x_{t-1}|x_t, x_0) = \sqrt{\bar{\alpha}_{t-1}}\, x_0 + \sqrt{1-\bar{\alpha}_{t-1}}\, \varepsilon \qquad \varepsilon \sim N(0, I)$$

$$= \sqrt{\bar{\alpha}_{t-1}}\, \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\, \varepsilon_\theta(x_t, t)\right) + \sqrt{1-\bar{\alpha}_{t-1}}\, \varepsilon$$

没有使用到 $q(x_t|x_{t-1})$

$$= \sqrt{\bar{\alpha}_{t-1}}\, \hat{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\, \varepsilon$$

$$= \sqrt{\bar{\alpha}_{t-1}}\, \hat{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}}\, \varepsilon_\theta(x_t, t)$$

$$= \sqrt{\bar{\alpha}_{t-1}}\, \hat{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\, \varepsilon_\theta + \sigma_t \varepsilon$$

$$q(x_s \mid x_k, x_0) = \sqrt{\bar{\alpha}_s}\, x_0 + \sqrt{1-\bar{\alpha}_s}\, \varepsilon \qquad \varepsilon \sim N(0, I)$$

$$= \sqrt{\bar{\alpha}_s}\, \frac{1}{\sqrt{\bar{\alpha}_k}} \left(x_k - \sqrt{1-\bar{\alpha}_k}\, \varepsilon_\theta(x_k, k) \right) + \sqrt{1-\bar{\alpha}_s}\, \varepsilon$$

$q(x_t \mid x_{t-1})$

$$= \sqrt{\bar{\alpha}_s}\, \hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s}\, \varepsilon$$

$$= \sqrt{\bar{\alpha}_s}\, \hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s}\, \varepsilon_\theta(x_k, k)$$

$$= \sqrt{\bar{\alpha}_s}\, \hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s - \sigma_k^2}\, \varepsilon_\theta + \sigma_k \varepsilon$$

$q(x_s \mid x_k, x_0)$

$$x_s = \sqrt{\bar{\alpha}_s}\, \hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s - \sigma_k^2}\, \varepsilon_\theta(x_k, k) + \sigma_k \varepsilon$$

$$x_s = \sqrt{\bar{\alpha}_s}\,\hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s-\sigma_k^2}\,\varepsilon_\theta(x_k, k) + \sigma_k \varepsilon$$

T=1000 999 998 ⋯ 3 2 1 0 (DDPM)

T=1000 ⋯ 900 888 666 233 ⋯ 0

一个符号差异

DDPM:

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

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

DDIM:

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

回顾DDPM

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \text{where } q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}\left(\sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\boldsymbol{x}_{t-1}, \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right)\boldsymbol{I}\right)$$

前向

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) := \int q(\boldsymbol{x}_{1:t}|\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_{1:(t-1)} = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I});$$

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

$$p_\theta(\boldsymbol{x}_0) = \int p_\theta(\boldsymbol{x}_{0:T})\mathrm{d}\boldsymbol{x}_{1:T}, \quad \text{where} \quad p_\theta(\boldsymbol{x}_{0:T}) := p_\theta(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \tag{1}$$

逆向

$$\max_\theta \mathbb{E}_{q(\boldsymbol{x}_0)}[\log p_\theta(\boldsymbol{x}_0)] \leq \max_\theta \mathbb{E}_{q(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)}[\log p_\theta(\boldsymbol{x}_{0:T}) - \log q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)]$$

$$L_\gamma(\epsilon_\theta) := \sum_{t=1}^{T} \gamma_t \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})}\left[\left\|\epsilon_\theta^{(t)}(\sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon_t) - \epsilon_t\right\|_2^2\right]$$

DDIM的思想

$$q(x_{t-1}|x_t) = \frac{q(x_t|x_{t-1})q(x_{t-1})}{q(x_t)}$$

$$\|$$

逆向过程 ☆

$$\boxed{q(x_{t-1}|x_t,x_0) = \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)} = N(x_{t-1};\mu(x_t,x_0),\sigma_t^2 I)}$$

$$q(x_t|x_{t-1}) = N(x_t;\sqrt{\alpha_t}x_{t-1},(1-\alpha_t)I)$$

$$x_0 = \frac{1}{\sqrt{\bar\alpha_t}}(x_t - \sqrt{1-\bar\alpha_t}\,\varepsilon_\theta)$$

$$\mu(x_t,x_0) = \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}x_t + \frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}x_0$$

$$\sigma_t^2 = \frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\cdot\beta_t$$

贝叶斯公式+两个边缘分布

$$\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) &= q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{x}_0)\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \\
&\propto \exp\Big(-\frac{1}{2}\big(\frac{(\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_{t-1})^2}{\beta_t}+\frac{(\mathbf{x}_{t-1}-\sqrt{\bar\alpha_{t-1}}\mathbf{x}_0)^2}{1-\bar\alpha_{t-1}}-\frac{(\mathbf{x}_t-\sqrt{\bar\alpha_t}\mathbf{x}_0)^2}{1-\bar\alpha_t}\big)\Big) \\
&= \exp\Big(-\frac{1}{2}\big((\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar\alpha_{t-1}})\mathbf{x}_{t-1}^2-(\frac{2\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{2\sqrt{\bar\alpha_t}}{1-\bar\alpha_t}\mathbf{x}_0)\mathbf{x}_{t-1}+C(\mathbf{x}_t,\mathbf{x}_0)\big)\Big)
\end{aligned}$$

在推导出L_{simple}（或L_γ）过程中，我们没有用到$q(x_{1:T}|x_0)$的具体形式，只是基于贝叶斯公式和$q(x_t|x_{t-1},x_0)$、$q(x_t|x_0)$表达式。

在训练DDPM所用到的L_{simple} loss中，我们甚至都没有采用跟$q(x_t|x_{t-1},x_0)$相关的系数，而是直接选择将预测噪音的权重设置为1。由于噪音项是来自$q(x_t|x_0)$的采样，因此，DDPM的目标函数其实只由$q(x_t|x_0)$表达式决定。这其实也证实了L_{simple}与score-matching之间的联系。

换句话说，只要$q(x_t|x_0)$已知并且是高斯分布的形式，那么我们就可以用DDPM的预测噪音的目标函数L_{simple}来训练模型。

于是，槽点来了！！！在DDPM中，基于马尔科夫性质我们认为$q(x_t|x_{t-1},x_0)$=$q(x_t|x_{t-1})$，那么如果是服从非马尔科夫性质呢？$q(x_t|x_{t-1},x_0)$是不是具有更一般的形式？以及，只要我们保证$q(x_t|x_0)$的形式不变，那么我们可以直接复用训练好的DDPM，只不过使用新的概率分布来进行逆过程的采样。

接下来作者给出了一种非马尔科夫性质的前向扩散过程的公式以及后验概率分布的表达式，而该后验概率分布恰好满足DDPM中的边缘分布$q(x_t|x_0)$。

作者给出了一种满足条件非马尔科夫过程

DDIM

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

$$q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_T}\boldsymbol{x}_0, (1-\alpha_T)\boldsymbol{I})$$

DDPM

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

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar\alpha_t}\mathbf{x}_0, (1-\bar\alpha_t)\mathbf{I})$$

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

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

$$\text{where}\quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) := \frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\mathbf{x}_t \quad\text{and}\quad \tilde\beta_t := \frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\beta_t$$

DDIM

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0, (1-\alpha_{t-1})\boldsymbol{I})$$

DDPM

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar\alpha_t}\mathbf{x}_0, (1-\bar\alpha_t)\mathbf{I})$$

B PROOFS

Lemma 1. *For $q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ defined in Eq. (6) and $q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ defined in Eq. (7), we have:*

$$q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I}) \tag{22}$$

Proof. Assume for any $t \leq T$, $q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I})$ holds, if:

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0, (1-\alpha_{t-1})\boldsymbol{I}) \tag{23}$$

then we can prove the statement with an induction argument for t from T to 1, since the base case ($t = T$) already holds.

First, we have that

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) := \int_{\boldsymbol{x}_t} q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \mathrm{d}\boldsymbol{x}_t$$

and

$$q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I}) \tag{24}$$

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 \boldsymbol{I}\right). \tag{25}$$

From Bishop (2006) (2.115), we have that $q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)$ is Gaussian, denoted as $\mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$ where

$$\begin{aligned} \mu_{t-1} &= \sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{\sqrt{\alpha_t}\boldsymbol{x}_0 - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}} &\quad (26)\\ &= \sqrt{\alpha_{t-1}}\boldsymbol{x}_0 &\quad (27) \end{aligned}$$

and

$$\Sigma_{t-1} = \sigma_t^2 \boldsymbol{I} + \frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}(1-\alpha_t)\boldsymbol{I} = (1-\alpha_{t-1})\boldsymbol{I} \tag{28}$$

Therefore, $q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0, (1-\alpha_{t-1})\boldsymbol{I})$, which allows us to apply the induction argument. $\square$

DDIM逆向过程

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{4}$$

For some $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $\boldsymbol{x}_t$ can be obtained using Eq. (4). The model $\epsilon_\theta^{(t)}(\boldsymbol{x}_t)$ then attempts to predict ϵ_t from $\boldsymbol{x}_t$, without knowledge of $\boldsymbol{x}_0$. By rewriting Eq. (4), one can then predict the *denoised observation*, which is a prediction of $\boldsymbol{x}_0$ given $\boldsymbol{x}_t$:

$$f_\theta^{(t)}(\boldsymbol{x}_t) := (\boldsymbol{x}_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t))/\sqrt{\alpha_t}. \tag{9}$$

We can then define the generative process with a fixed prior $p_\theta(\boldsymbol{x}_T) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and

DDIM

$$p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(\boldsymbol{x}_1), \sigma_1^2\boldsymbol{I}) & \text{if } t = 1 \\ q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_\theta^{(t)}(\boldsymbol{x}_t)) & \text{otherwise,} \end{cases} \tag{10}$$

DDPM

$$\underbrace{q(x_{t-1}|x_t)}_{\parallel} = \frac{q(x_t|x_{t-1})\,q(x_{t-1})}{q(x_t)}$$

☆

$$\boxed{q(x_{t-1}|x_t, x_0) = \frac{\boxed{q(x_t|x_{t-1})}\,q(x_{t-1}|x_0)}{q(x_t|x_0)} = N(x_{t-1}; \mu(x_t, x_0), \sigma_t^2 I)}$$

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}\,x_{t-1}, (1-\alpha_t)I)$$

$$\hat{x}_{0|t} = \frac{1}{\sqrt{\bar\alpha_t}}(x_t - \sqrt{1-\bar\alpha_t}\,\varepsilon_\theta)$$

损失函数

DDIM

We optimize θ via the following variational inference objective (which is a functional over ϵ_θ):

$$J_\sigma(\epsilon_\theta) := \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})}[\log q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) - \log p_\theta(\boldsymbol{x}_{0:T})] \tag{11}$$

$$= \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})}\left[\log q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) + \sum_{t=2}^{T} \log q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) - \log p_\theta(\boldsymbol{x}_T)\right]$$

where we factorize $q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ according to Eq. (6) and $p_\theta(\boldsymbol{x}_{0:T})$ according to Eq. (1).

DDPM

$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + C \tag{8}$$

where C is a constant that does not depend on θ. So, we see that the most straightforward parameterization of $\boldsymbol{\mu}_\theta$ is a model that predicts $\tilde{\boldsymbol{\mu}}_t$, the forward process posterior mean. However, we can expand Eq. (8) further by reparameterizing Eq. (4) as $\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ for $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and applying the forward process posterior formula (7):

$$L_{t-1} - C = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon})\right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), t)\right\|^2\right] \tag{9}$$

$$= \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}}\left[\frac{1}{2\sigma_t^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}) - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) - \boldsymbol{\mu}_\theta(\mathbf{x}_t(\mathbf{x}_0, \boldsymbol{\epsilon}), t)\right\|^2\right] \tag{10}$$

Theorem 1. *For all $\sigma > \mathbf{0}$, there exists $\gamma \in \mathbb{R}^T_{>0}$ and $C \in \mathbb{R}$, such that $J_\sigma = L_\gamma + C$.*

Proof. From the definition of J_σ:

$$J_\sigma(\epsilon_\theta) := \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q(\boldsymbol{x}_{0:T})}\left[\log q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) + \sum_{t=2}^{T} \log q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\right] \tag{29}$$

$$\equiv \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q(\boldsymbol{x}_{0:T})}\left[\sum_{t=2}^{T} D_{\mathrm{KL}}(q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0))\|p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) - \log p_\theta^{(1)}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]$$

where we use $\equiv$ to denote "equal up to a value that does not depend on ϵ_θ (but may depend on q_σ)". For $t > 1$:

$$\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)}[D_{\mathrm{KL}}(q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0))\|p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))]$$

$$= \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)}[D_{\mathrm{KL}}(q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0))\|q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_\theta^{(t)}(\boldsymbol{x}_t)))]$$

$$\equiv \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)}\left[\frac{\|\boldsymbol{x}_0 - f_\theta^{(t)}(\boldsymbol{x}_t)\|_2^2}{2\sigma_t^2}\right] \tag{30}$$

$$= \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon}\left[\frac{\left\|\frac{(\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon)}{\sqrt{\alpha_t}} - \frac{(\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t))}{\sqrt{\alpha_t}}\right\|_2^2}{2\sigma_t^2}\right] \tag{31}$$

$$= \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon}\left[\frac{\|\epsilon - \epsilon_\theta^{(t)}(\boldsymbol{x}_t)\|_2^2}{2d\sigma_t^2\alpha_t}\right] \tag{32}$$

where d is the dimension of $\boldsymbol{x}_0$. For $t = 1$:

$$\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_1 \sim q(\boldsymbol{x}_0, \boldsymbol{x}_1)}\left[-\log p_\theta^{(1)}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] \equiv \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_1 \sim q(\boldsymbol{x}_0, \boldsymbol{x}_1)}\left[\frac{\|\boldsymbol{x}_0 - f_\theta^{(t)}(\boldsymbol{x}_1)\|_2^2}{2\sigma_1^2}\right] \tag{33}$$

$$= \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \boldsymbol{x}_1 = \sqrt{\alpha_1}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon}\left[\frac{\|\epsilon - \epsilon_\theta^{(1)}(\boldsymbol{x}_1)\|_2^2}{2d\sigma_1^2\alpha_1}\right] \tag{34}$$

Therefore, when $\gamma_t = 1/(2d\sigma_t^2\alpha_t)$ for all $t \in \{1, \dots, T\}$, we have

$$J_\sigma(\epsilon_\theta) \equiv \sum_{t=1}^{T} \frac{1}{2d\sigma_t^2\alpha_t}\mathbb{E}\left[\|\epsilon_\theta^{(t)}(\boldsymbol{x}_t) - \epsilon_t\|_2^2\right] = L_\gamma(\epsilon_\theta) \tag{35}$$

for all ϵ_θ. From the definition of "$\equiv$", we have that $J_\sigma = L_\gamma + C$. $\square$

逆向过程采样

$$p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(\boldsymbol{x}_1), \sigma_1^2 \boldsymbol{I}) & \text{if } t = 1 \\ q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_\theta^{(t)}(\boldsymbol{x}_t)) & \text{otherwise,} \end{cases}$$

DDIM

$$f_\theta^{(t)}(\boldsymbol{x}_t) := (\boldsymbol{x}_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t))/\sqrt{\alpha_t}.$$

From $p_\theta(\boldsymbol{x}_{1:T})$ in Eq. (10), one can generate a sample $\boldsymbol{x}_{t-1}$ from a sample $\boldsymbol{x}_t$ via:

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \right)}_{\text{" predicted } \boldsymbol{x}_0 \text{"}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{"direction pointing to } \boldsymbol{x}_t \text{"}} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}} \tag{12}$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ is standard Gaussian noise independent of $\boldsymbol{x}_t$, and we define $\alpha_0 := 1$. Different choices of σ values results in different generative processes, all while using the same model ϵ_θ, so re-training the model is unnecessary. When $\sigma_t = \sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$ for all t, the forward process becomes Markovian, and the generative process becomes a DDPM.

DDPM

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

$$q(x_s \mid x_k, x_0) = \sqrt{\bar{\alpha}_s}\, x_0 + \sqrt{1-\bar{\alpha}_s}\, \varepsilon \qquad \varepsilon \sim N(0, I)$$

$$= \sqrt{\bar{\alpha}_s}\, \frac{1}{\sqrt{\bar{\alpha}_k}} \left(x_k - \sqrt{1-\bar{\alpha}_k}\, \varepsilon_\theta(x_k, k) \right) + \sqrt{1-\bar{\alpha}_s}\, \varepsilon$$

$q(x_t \mid x_{t-1})$

$$= \sqrt{\bar{\alpha}_s}\, \hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s}\, \varepsilon$$

$$= \sqrt{\bar{\alpha}_s}\, \hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s}\, \varepsilon_\theta(x_k, k)$$

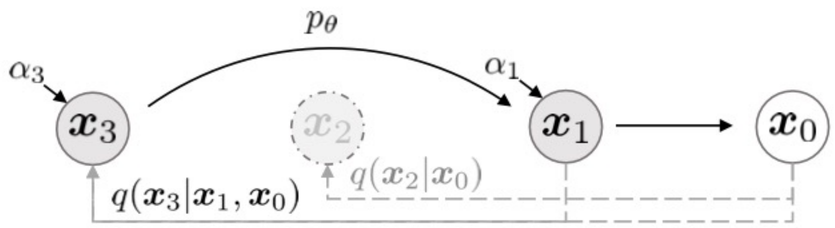
$$= \sqrt{\bar{\alpha}_s}\, \hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s-\sigma_k^2}\, \varepsilon_\theta + \sigma_k \varepsilon$$

$q(x_s \mid x_k, x_0)$

$$x_s = \sqrt{\bar{\alpha}_s}\, \hat{x}_{0|k} + \sqrt{1-\bar{\alpha}_s-\sigma_k^2}\, \varepsilon_\theta(x_k, k) + \sigma_k \varepsilon$$

T=1000 999 998 ⋯ 3 2 1 0 (DDPM)

T=1000 ⋯ 900 888 666 233 ⋯ 0



DDPM: $\bar{\alpha}_1 \quad \bar{\alpha}_2 \quad \cdots \quad \bar{\alpha}_t \quad \cdots \quad \bar{\alpha}_{T-1} \quad \bar{\alpha}_T$

$\downarrow$

$x_0 \to x_t$

$\downarrow$

$\epsilon_\theta(x_1, 1) \quad \epsilon_\theta(x_2, 2) \quad \epsilon_\theta(x_t, t) \quad \cdots \quad \epsilon_\theta(x_T, T)$

DDIM $\bar{\alpha}_{100} \quad \bar{\alpha}_{200} \quad \cdots \quad \bar{\alpha}_{900} \quad \bar{\alpha}_{1000} = \bar{\alpha}_T$

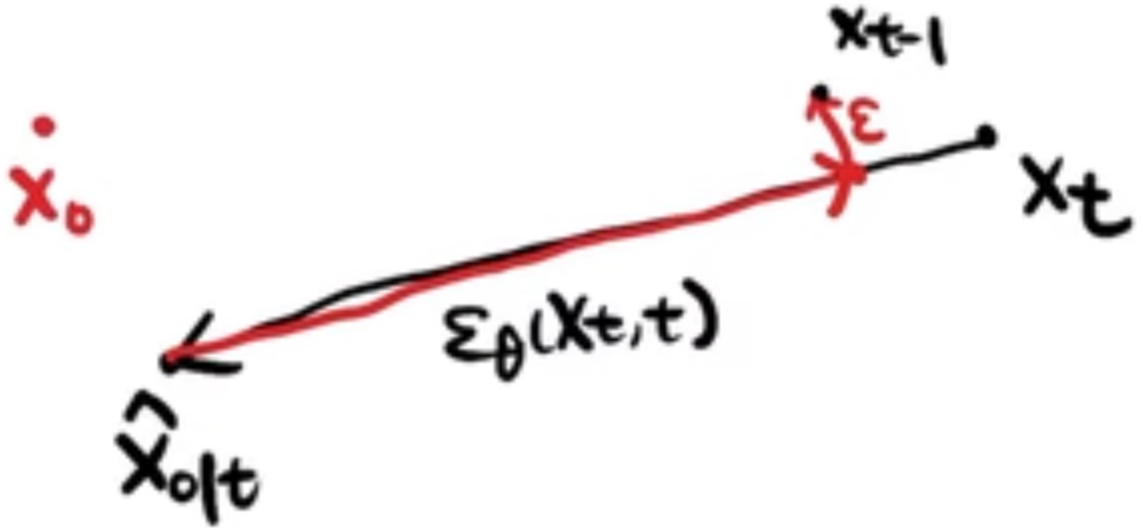
$\epsilon_\theta(x_{100}, 100) \quad \epsilon_\theta(x_{200}, 200) \quad \cdots \quad \epsilon_\theta(x_T, T)$

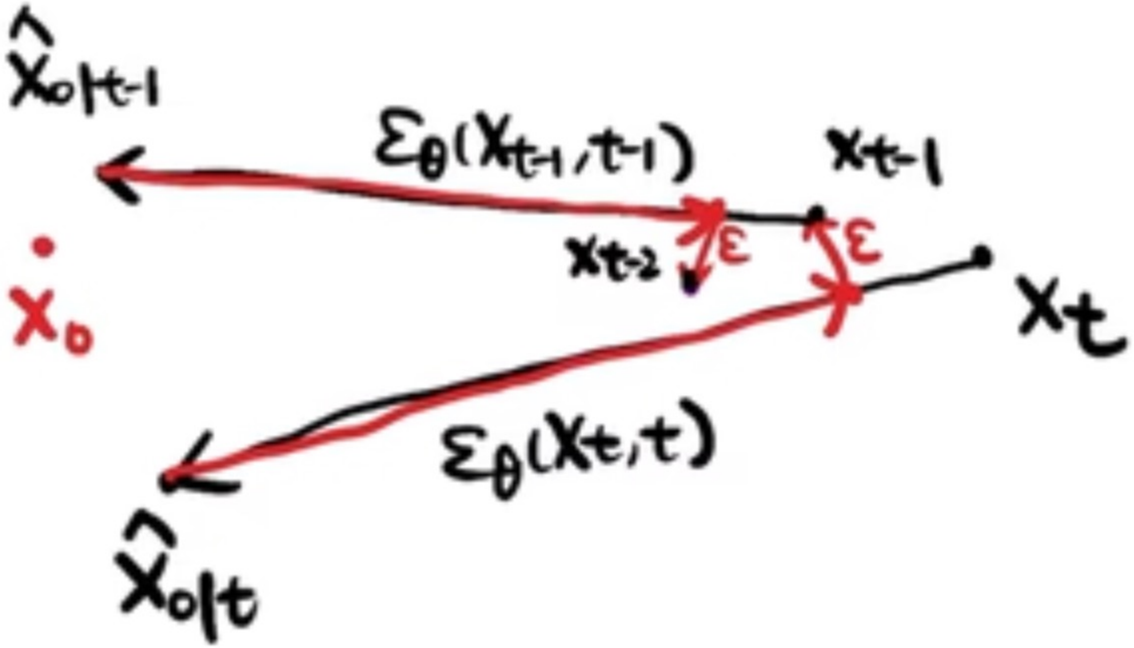
$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}).$$

图解

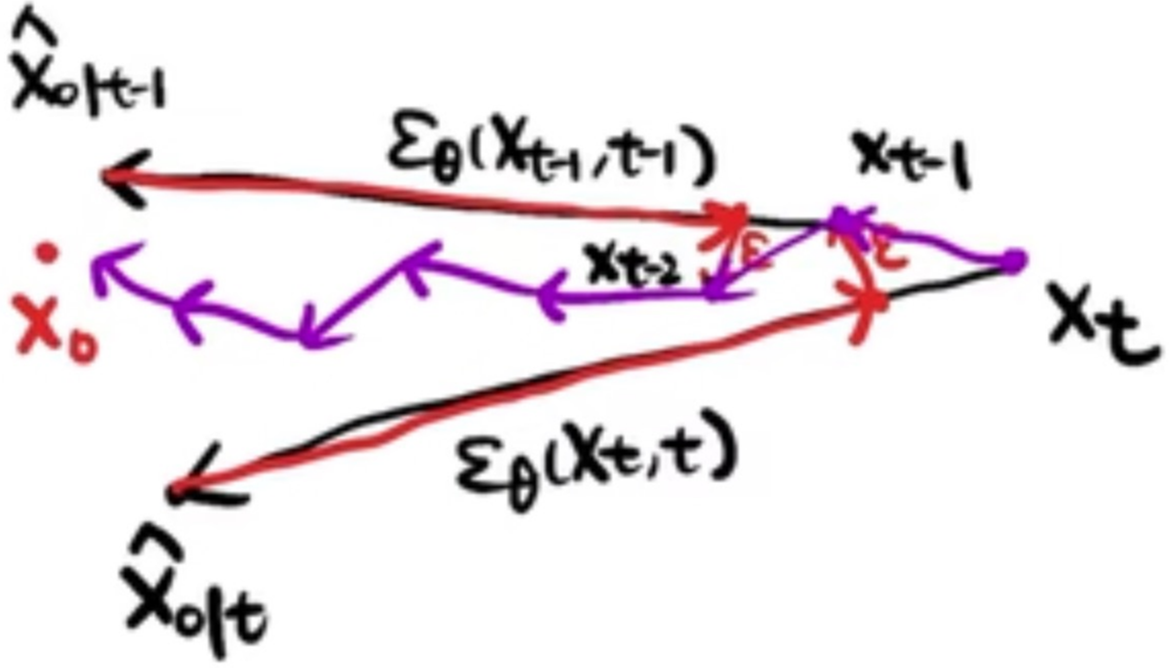
$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}\right)}_{\text{“ predicted } \boldsymbol{x}_0\text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{“direction pointing to } \boldsymbol{x}_t\text{”}} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}}$$

$$x_{t-1} = \underline{\sqrt{\bar\alpha_{t-1}}\,\hat{x}_{0|t}} + \underline{\sqrt{1-\bar\alpha_{t-1}-\sigma_t^2}\;\varepsilon_\theta(x_t,t)} + \sigma_t\varepsilon$$





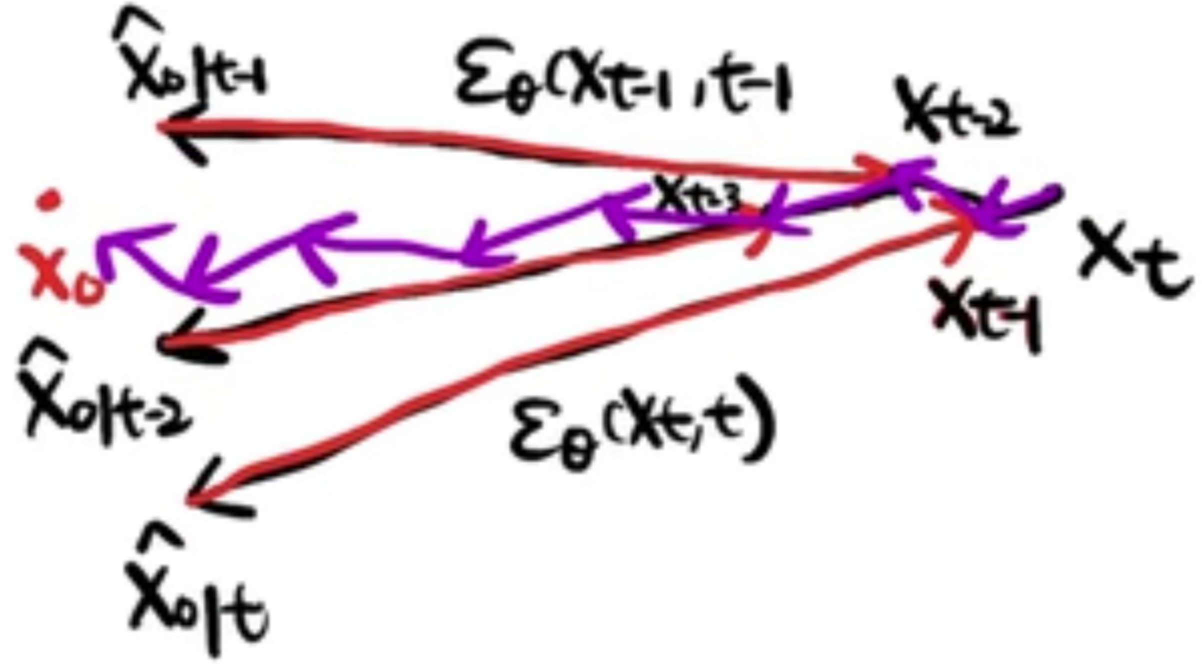
$\hat{x}_{0|t-1}$
$\epsilon_\theta(x_{t-1}, t-1)$
x_{t-1}
x_{t-2}
ϵ
ϵ
x_0
x_t
$\epsilon_\theta(x_t, t)$
$\hat{x}_{0|t}$



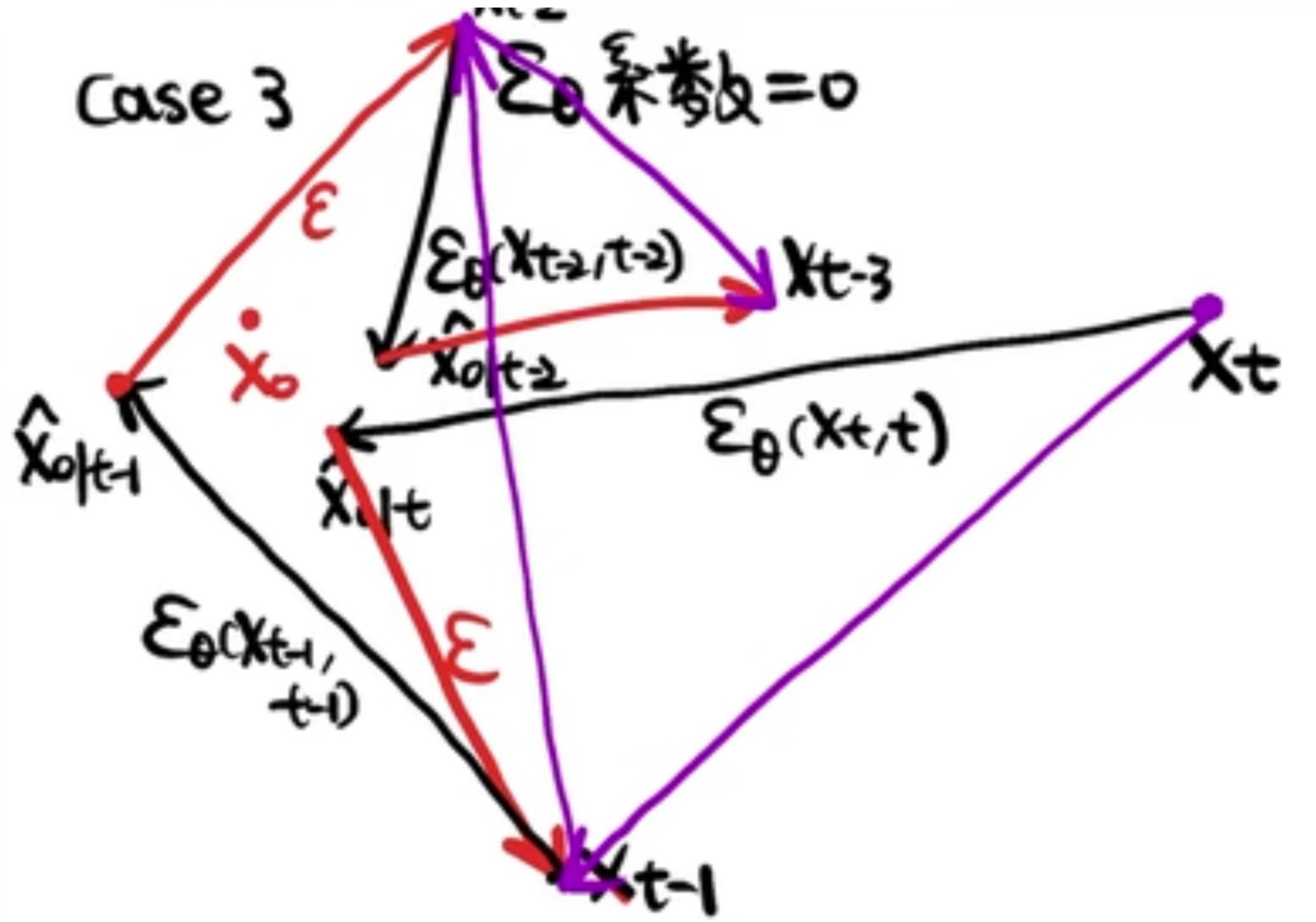
$\hat{x}_{0|t-1}$
$\epsilon_\theta(x_{t-1}, t-1)$
x_{t-1}
x_{t-2}
ϵ
ϵ
$\dot{x}_0$
x_t
$\epsilon_\theta(x_t, t)$
$\hat{x}_{0|t}$

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\;\varepsilon_\theta(x_t, t) + \sigma_t \varepsilon$$

case2 $\sigma_t = 0$



$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\epsilon_\theta(x_t,t) + \sigma_t\epsilon$$



常微分方程

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\, \epsilon_\theta(x_t, t) + \sigma_t \epsilon$$

$\sigma_t = 0$ $\quad x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, x_0 + \sqrt{1-\bar{\alpha}_{t-1}}\, \epsilon_\theta(x_t, t)$

$$= \sqrt{\bar{\alpha}_{t-1}}\, \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \sqrt{1-\bar{\alpha}_t}\, \epsilon_\theta(x_t, t)\right) + \sqrt{1-\bar{\alpha}_{t-1}}\, \epsilon_\theta(x_t, t)$$

$$\frac{x_{t-1}}{\sqrt{\bar{\alpha}_{t-1}}} = \frac{x_t}{\sqrt{\bar{\alpha}_t}} - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\, \epsilon_\theta(x_t, t) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_{t-1}}}\, \epsilon_\theta(x_t, t)$$

$$\frac{x_t}{\sqrt{\bar{\alpha}_t}} - \frac{x_{t-1}}{\sqrt{\bar{\alpha}_{t-1}}} = \left(\sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}} - \sqrt{\frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_{t-1}}}\right) \epsilon_\theta(x_t, t)$$

$a = \sqrt{\bar{\alpha}}$

$\sigma = \sqrt{\frac{1-\bar{\alpha}}{\bar{\alpha}}}$

$$\frac{d}{ds}\left(\frac{x(s)}{a(s)}\right) = \frac{d}{ds}\sigma(s)\, \epsilon_\theta(x(s), t(s))$$

$s \in [0,1]$ 已知 $x(1) \sim N(0, I)$ 求 $x(0)$

ODE form for VE-SDE Define $p_t(\bar{\boldsymbol{x}})$ as the data distribution perturbed with $\sigma^2(t)$ variance Gaussian noise. The probability flow for VE-SDE is defined as Song et al. (2020):

$$\mathrm{d}\bar{\boldsymbol{x}} = -\frac{1}{2} g(t)^2 \nabla_{\bar{\boldsymbol{x}}} \log p_t(\bar{\boldsymbol{x}}) \mathrm{d}t \tag{47}$$

where $g(t) = \sqrt{\frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t}}$ is the diffusion coefficient, and $\nabla_{\bar{\boldsymbol{x}}} \log p_t(\bar{\boldsymbol{x}})$ is the score of p_t.

The $\sigma(t)$-perturbed score function $\nabla_{\bar{\boldsymbol{x}}} \log p_t(\bar{\boldsymbol{x}})$ is also a minimizer (from denoising score matching (Vincent, 2011)):

$$\nabla_{\bar{\boldsymbol{x}}} \log p_t = \arg\min_{g_t} \mathbb{E}_{\boldsymbol{x}(0) \sim q(\boldsymbol{x}), \epsilon \sim \mathcal{N}(0, \boldsymbol{I})} [\| g_t(\bar{\boldsymbol{x}}) + \epsilon / \sigma(t) \|_2^2] \tag{48}$$

where $\bar{\boldsymbol{x}}(t) = \bar{\boldsymbol{x}}(t) + \sigma(t)\epsilon$.

Since there is an equivalence between $\boldsymbol{x}(t)$ and $\bar{\boldsymbol{x}}(t)$, we have the following relationship:

$$\nabla_{\bar{\boldsymbol{x}}} \log p_t(\bar{\boldsymbol{x}}) = -\frac{\epsilon_\theta^{(t)}\left(\frac{\bar{\boldsymbol{x}}(t)}{\sqrt{\sigma^2(t)+1}}\right)}{\sigma(t)} \tag{49}$$

from Equation (46) and Equation (48). Plug Equation (49) and definition of $g(t)$ in Equation (47), we have:

$$\mathrm{d}\bar{\boldsymbol{x}}(t) = \frac{1}{2} \frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t} \frac{\epsilon_\theta^{(t)}\left(\frac{\bar{\boldsymbol{x}}(t)}{\sqrt{\sigma^2(t)+1}}\right)}{\sigma(t)} \mathrm{d}t, \tag{50}$$

and we have the following by rearranging terms:

$$\frac{\mathrm{d}\bar{\boldsymbol{x}}(t)}{\mathrm{d}t} = \frac{\mathrm{d}\sigma(t)}{\mathrm{d}t} \epsilon_\theta^{(t)} \left(\frac{\bar{\boldsymbol{x}}(t)}{\sqrt{\sigma^2(t)+1}}\right) \tag{51}$$

which is equivalent to Equation (45). In both cases the initial conditions are $\bar{\boldsymbol{x}}(T) \sim \mathcal{N}(\mathbf{0}, \sigma^2(T)\boldsymbol{I})$, so the resulting ODEs are identical. □