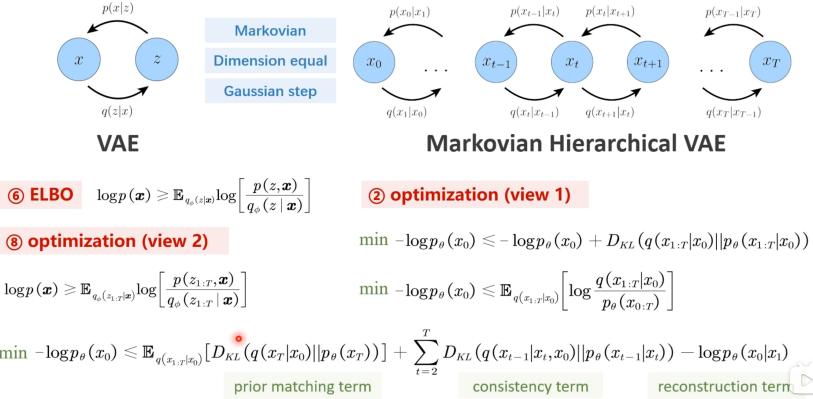
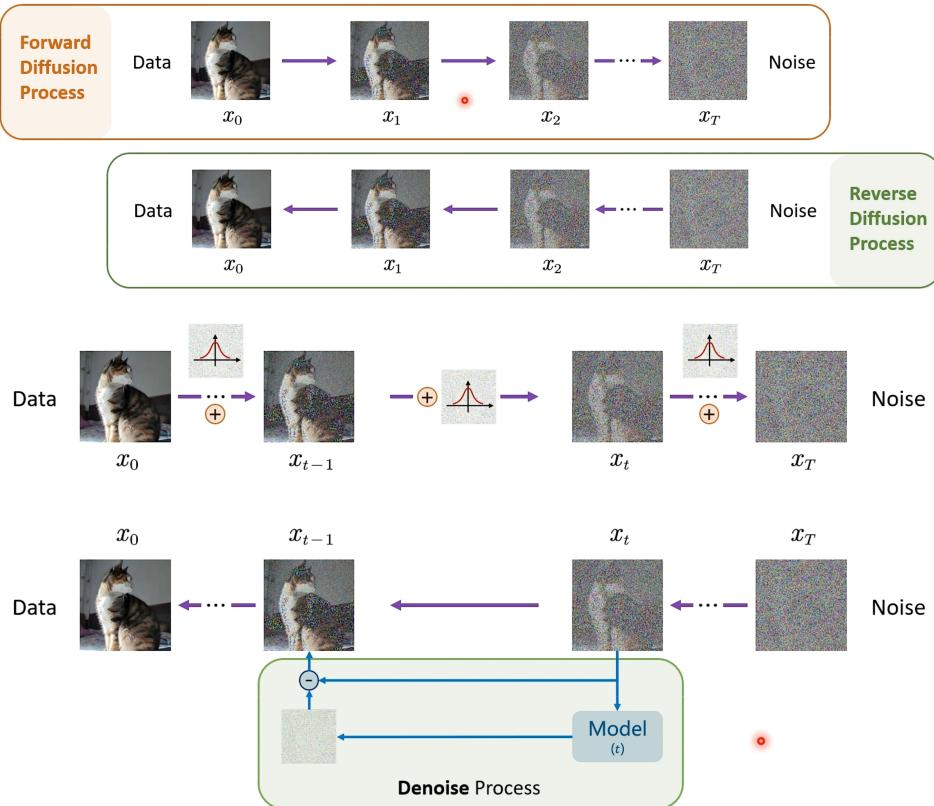


Diffusion Model

Relationship with diffusion



Denoising Diffusion Probabilistic Models^[1]



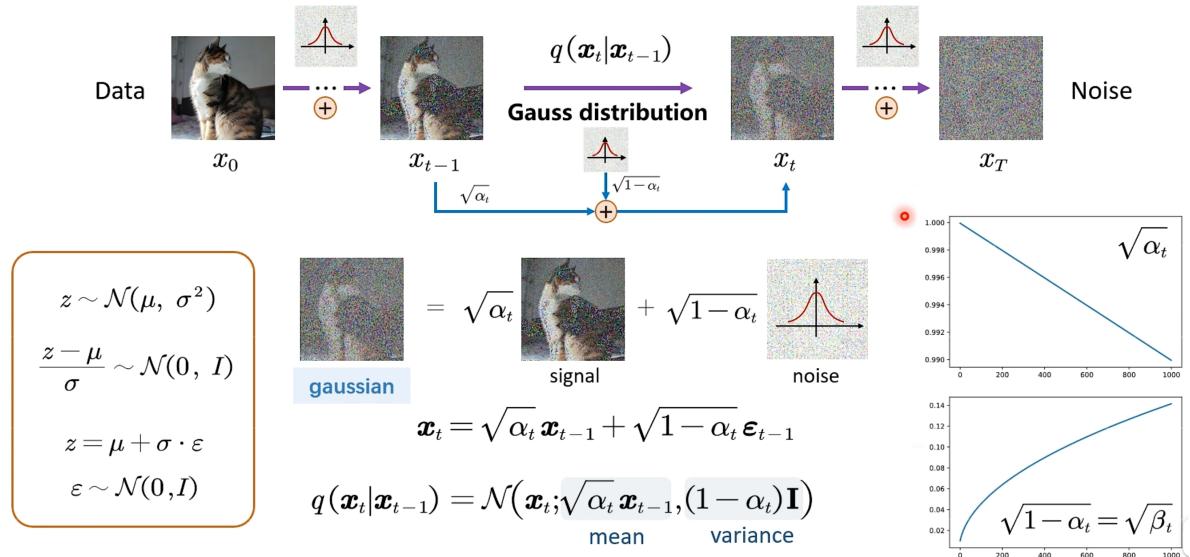
Forward Process

输入样本 x_0 通过加 T 次噪声达到 x_T , $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 。Markov 过程被定义为:

$$x_t \sim \mathcal{N}(\mu_t x_{t-1}, \sigma_t^2 \mathbf{I}) = \mu_t x_{t-1} + \epsilon \odot \sigma_t \mathbf{I}$$

Diffusion Noise 过程被定义为:

$$\begin{aligned} \begin{cases} \mu_t = \sqrt{\alpha_t} \\ \sigma_t = \sqrt{1 - \alpha_t} \end{cases} \Rightarrow \begin{cases} \mu_t = \sqrt{1 - \beta_t} \\ \sigma_t = \sqrt{\beta_t} \end{cases} \\ x_t \sim \mathcal{N}(\sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}) \Rightarrow x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon_{t-1} \\ = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon' \end{aligned}$$



我们使用一个加噪音 Dist , 其被定义为:

$$x_t = \underbrace{\sqrt{\alpha_t} x_{t-1}}_{\mu} + \underbrace{\sqrt{1 - \alpha_t}}_{\sigma} \varepsilon_{t-1}, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

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

通过多次迭代, 可发现:

$$\begin{aligned} x_t &= \sqrt{\alpha_t} x_{t-1} + \sqrt{1 - \alpha_t} \varepsilon_{t-1} \\ &= \sqrt{\alpha_t} (\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1 - \alpha_{t-1}} \varepsilon_{t-2}) + \sqrt{1 - \alpha_t} \varepsilon_{t-1} \\ &= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \underbrace{(\sqrt{\alpha_t(1 - \alpha_t)} \varepsilon_{t-2} + \sqrt{1 - \alpha_t} \varepsilon_{t-1})}_{\sqrt{\alpha_t(1 - \alpha_t)} + (1 - \alpha_t) \varepsilon'} \end{aligned}$$

$$\begin{aligned}
 &= \dots \\
 &= \sqrt{\alpha_t \alpha_{t-1} \dots \alpha_1} x_0 + \underbrace{\sqrt{1 - \alpha_t \alpha_{t-1} \dots \alpha_1}}_{\Gamma_\varepsilon} \varepsilon' \\
 &= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon' \\
 \bar{\alpha}_t &\triangleq \alpha_t \alpha_{t-1} \dots \alpha_1
 \end{aligned}$$

对于两个高斯分布 $X \sim \mathcal{N}(\mu_X, \sigma_X^2), Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. 若 $Z = X + Y$, 则有:

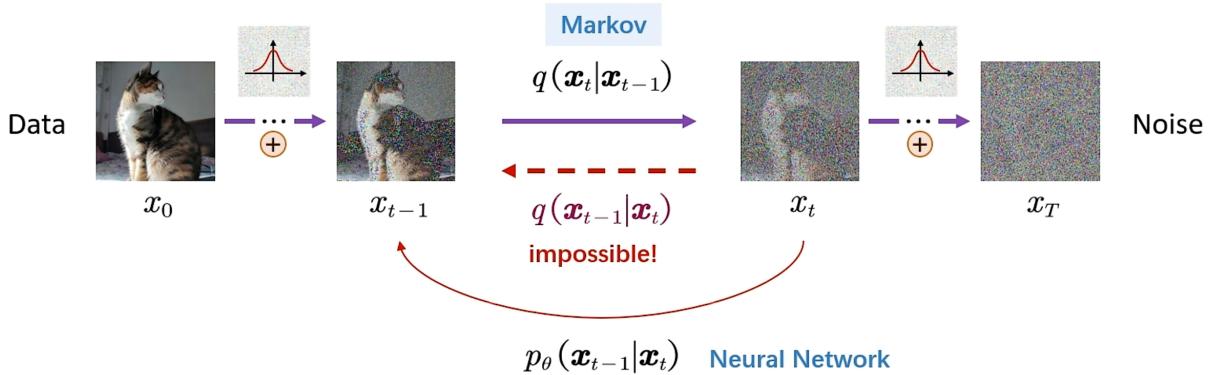
$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Γ_ε 如下:

$$\begin{aligned}
 \varepsilon_{t-1} &\Rightarrow \sqrt{1 - \alpha_t} \\
 \varepsilon_{t-2} &\Rightarrow \sqrt{\alpha_t(1 - \alpha_{t-1})} \\
 \varepsilon_{t-2} &\Rightarrow \sqrt{\alpha_t \alpha_{t-1}(1 - \alpha_{t-2})} \\
 &\dots \\
 \varepsilon_0 &\Rightarrow \sqrt{\alpha_t \alpha_{t-1} \dots \alpha_2(1 - \alpha_1)}
 \end{aligned}$$

$$\begin{aligned}
 S &= (1 - \alpha_t) + [\alpha_t(1 - \alpha_{t-1})] + [\alpha_t \alpha_{t-1}(1 - \alpha_{t-2})] + \dots + [\alpha_t \alpha_{t-1} \dots \alpha_2(1 - \alpha_1)] \\
 &= 1 - \cancel{\alpha_t} + \cancel{\alpha_t} - \cancel{\alpha_t \alpha_{t-1}} + \cancel{\alpha_t \alpha_{t-1}} - \cancel{\alpha_t \alpha_{t-1} \alpha_{t-2}} + \dots + \cancel{\alpha_t \alpha_{t-1} \dots \alpha_2} - \cancel{\alpha_t \alpha_{t-1} \dots \alpha_2 \alpha_1} \\
 &= 1 - \alpha_t \alpha_{t-1} \dots \alpha_2 \alpha_1
 \end{aligned}$$

Reverse Process



当 β_t (i.e., $1 - \alpha_t$) 足够小时, Noise 的逆操作也符合正态分布, 也就是:

$$x_{t-1} \sim \mathcal{N}(\tilde{\mu}_t, \tilde{\sigma}_t^2 \mathbf{I})$$

这里的 $\tilde{\mu}_t, \tilde{\sigma}_t^2$ 是由当前时刻 t 和图像 x_t 决定的, 因此我们需要 NN 去拟合:

$$\tilde{\mu}_t, \tilde{\sigma}_t^2 = p_\theta(x_{t-1} \mid t, x_t)$$

直接寻找反向概率分布 $q(x_{t-1} \mid x_t)$ 是不可能的，使用 NN $p_\theta(x_{t-1} \mid x_t)$ 以拟合

目标分布 $q(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_t(x_t), \Sigma_t(x_t))$

近似分布 $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$

考虑 $p(x_t \mid x_{t-1})$ 是一个 Markov 过程，而 $q(x_{t-1} \mid x_t)$ 也是一个 Markov 过程，因此

$$q(x_{t-1} \mid x_t) = q(x_{t-1} \mid x_t, x_0) = \frac{q(x_t \mid x_{t-1}, x_0)q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)}$$

这个概率分布左侧是不可知的，但是右侧我们已知：

$$\begin{aligned} q(x_t \mid x_{t-1}, x_0) &= \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I}) \\ q(x_{t-1} \mid x_0) &= \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1 - \bar{\alpha}_{t-1})\mathbf{I}) \\ q(x_t \mid x_0) &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}) \end{aligned}$$

也就是右侧全都知道了。

$$\begin{aligned} \mathcal{N}(\tilde{\mu}_t, \tilde{\sigma}_t^2 \mathbf{I}) &= \frac{\mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I})\mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1 - \bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I})} \\ \log \mathcal{N}(\tilde{\mu}_t, \tilde{\sigma}_t^2 \mathbf{I}) &= \log \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I}) + \log \mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1 - \bar{\alpha}_{t-1})\mathbf{I}) \\ &\quad - \log \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}) \\ \log \mathcal{N}(x; \mu, \sigma^2) &= -\frac{1}{2\sigma^2}(x - \mu)^2 + C \end{aligned}$$

$$\begin{aligned} \log \mathcal{N}(\tilde{\mu}_t, \tilde{\sigma}_t^2 \mathbf{I}) &= -\frac{1}{2} \left\{ \left[\frac{1}{1 - \alpha_t} (x_t - \sqrt{\alpha_t}x_{t-1})^2 + C_1 \right] \right. \\ &\quad \left. + \left[\frac{1}{1 - \bar{\alpha}_{t-1}} (x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0)^2 + C_2 \right] - \left[\frac{1}{1 - \bar{\alpha}_t} (x_t - \sqrt{\bar{\alpha}_t}x_0)^2 + C_3 \right] \right\} \end{aligned}$$

可以求出

$$\begin{cases} \tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_t \right) \\ \tilde{\sigma}_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \cdot \beta_t \end{cases}$$

会发现 $\tilde{\sigma}_t^2$ 不依赖 x_t 是常量。模型去拟合 $\tilde{\mu}_t$ 即可。而 $\tilde{\mu}_t$ 中只有 ε_t 不知道，因此只需要拟合 ε_t 即可。即 $\epsilon_\theta(x_t, t) \rightarrow \varepsilon_t$ 。而误差可以定义为：

$$L = \|\varepsilon_t - \epsilon_\theta(x_t, t)\|_2^2$$

Algorithm 1 Training

```

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

```

Algorithm 2 Sampling

```

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

```

Mathematics Behind Diffusion

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

给定概率

$$p(x) = \mathcal{N}(x; \mu_x, \sigma_x^2) \quad p(y | x) = \mathcal{N}(y; ax, \sigma_y^2)$$

则 Marginal Distribution $p(y)$ 和 Posterior Distribution $p(x | y)$ 也是正态分布

$$\begin{aligned} p(y) &= \mathcal{N}(y; ax, a^2\sigma_x^2 + \sigma_y^2) \\ p(x | y) &= \mathcal{N}(x; \tilde{\mu}, \tilde{\sigma}^2), \quad \frac{1}{\tilde{\sigma}^2} = \frac{1}{\sigma_x^2} + \frac{a^2}{\sigma_y^2}, \tilde{\mu} = \tilde{\sigma}^2(a\sigma_y^{-1}y + \sigma_x^{-1}\mu_x) \\ p(x | y) &= \frac{p(y | x)p(x)}{p(y)} = \frac{\mathcal{N}(y; ax, \sigma_y^2)\mathcal{N}(x; \mu_x, \sigma_x^2)}{\mathcal{N}(y; ax, a^2\sigma_x^2 + \sigma_y^2)} \end{aligned}$$

考虑 $p(y)$ 是归一化系数，因此：

$$\begin{aligned} p(x | y) &\equiv p(x)p(y | x) \\ &\propto \exp\left[-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right] \exp\left[-\frac{1}{2}\left(\frac{y-ax}{\sigma_y}\right)^2\right] \\ &= \exp\left[-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right] \exp\left[-\frac{1}{2}\frac{(y-ax)^2}{\sigma_y^2}\right] \\ &= \exp\left[-\frac{1}{2}\left\{\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-ax)^2}{2\sigma_y^2}\right\}\right] \\ &= \exp\left[-\frac{1}{2}\left\{\frac{x^2 + \mu_x^2 - 2\mu_x x}{\sigma_x^2} + \frac{y^2 + a^2x^2 - 2axy}{\sigma_y^2}\right\}\right] \\ &= \exp\left[-\frac{1}{2}\left\{x^2 \left(\frac{1}{\sigma_x^2} + \frac{a^2}{\sigma_y^2}\right) - 2x \left(\frac{\mu_x}{\sigma_x^2} + \frac{ay}{\sigma_y^2}\right) + \left(\frac{\mu_x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right\}\right] \end{aligned}$$

对比系数可知

$$\begin{aligned} \left(\frac{x-\mu}{\sigma}\right)^2 &= \frac{x^2 - 2\mu x + \mu^2}{\sigma^2} = \frac{x^2}{\sigma^2} - \frac{2\mu x}{\sigma^2} + \frac{\mu^2}{\sigma^2} \\ \frac{1}{\sigma^2} &= \frac{1}{\sigma_x^2} + \frac{a^2}{\sigma_y^2} \implies \sigma^2 = \frac{\sigma_x^2 \sigma_y^2}{\sigma_y^2 + a\sigma_x^2} \\ \mu &= \underbrace{\left(\frac{\sigma_x^2 \sigma_y^2}{\sigma_y^2 + a\sigma_x^2}\right)}_{\sigma^2} \underbrace{\left(\frac{\mu_x}{\sigma_x^2} + \frac{ay}{\sigma_y^2}\right)}_{a\sigma_y^{-1}y + \sigma_x^{-1}\mu_x}^{\mu\sigma^{-2}} \implies \mu = \frac{\sigma_y^2 \mu_x + a\sigma_x^2 y}{\sigma_y^2 + a\sigma_x^2} \end{aligned}$$

$$p(x | y) = \mathcal{N} \left(x; \frac{\sigma_x^2 \sigma_y^2}{\sigma_y^2 + a\sigma_x}, \frac{\sigma_y^2 \mu_x + a\sigma_x^2 y}{\sigma_y^2 + a\sigma_x^2} \right)$$

Probabilistic Graphical Models (PGM)

考慮 DAG $x_{t-1} \rightarrow x_t \rightarrow x_{t+1}$:

$$\begin{aligned} p(x_{t-1}, x_{t+1} | x_t) &= \frac{p(x_{t-1}, x_t, x_{t+1})}{p(x_t)} = \frac{p(x_{t-1}, x_t)p(x_{t+1} | x_t, x_{t-1})}{p(x_t)} \\ &= p(x_{t-1} | x_t)p(x_{t+1} | x_t) \end{aligned}$$

Maximise Likelihood Estimate (MLE)

$$\begin{aligned} \theta^* &= \operatorname{argmax}_{\theta} \prod_{x \in \mathcal{X}} p_{\theta}(x) = \operatorname{argmax}_{\theta} \log \prod_{x \in \mathcal{X}} p_{\theta}(x) \\ &= \operatorname{argmax}_{\theta} \sum_{x \in \mathcal{X}} \log p_{\theta}(x) \approx \operatorname{argmax}_{\theta} \mathbb{E}_{x \sim p_{data}} [\log p_{\theta}(x)] \\ &= \operatorname{argmax}_{\theta} \int_x p_{data}(x) \log p_{\theta}(x) dx \\ &= \operatorname{argmax}_{\theta} \int_x p_{data}(x) \log p_{\theta}(x) dx - \int_x p_{data}(x) \log p_{data}(x) dx \\ &= \operatorname{argmax}_{\theta} \int_x p_{data}(x) \log \frac{p_{\theta}(x)}{p_{data}(x)} dx \\ &= \operatorname{argmax}_{\theta} -KL[p_{data}(x) \| p_{\theta}(x)] \\ &= \operatorname{argmin}_{\theta} KL[p_{data}(x) \| p_{\theta}(x)] \end{aligned}$$

VAE and Diffusion

VAE:

$$\begin{aligned} \log p_{\theta}(x) &= \int_z q(z | x) \log p_{\theta}(x) dz = \int_z q(z | x) \log \left(\frac{p_{\theta}(x, z)q(z | x)}{q(z | x)p(z)} \right) dz \\ &= \underbrace{\mathbb{E}_{q(z|x)} \left[\log \left(\frac{p_{\theta}(x, z)}{q(z | x)} \right) \right]}_{\mathcal{L}_{VAE}} + KL[q(z | x) \| p(z)] \\ &\quad \text{maxmise } \mathbb{E}_{q(z|x)} \left[\log \left(\frac{p_{\theta}(x, z)}{q(z | x)} \right) \right] \end{aligned}$$

DDPM:

$$\begin{aligned} \log p_{\theta}(x_0) &= \int_{x_1:x_T} p(x_T)p_{\theta}(x_{T-1} | x_T) \cdots p_{\theta}(x_{t-1} | x_t) \cdots p_{\theta}(x_0 | x_1) dx_1: x_T \\ &\quad \text{maxmise } \mathbb{E}_{q(x_1:x_T|x_0)} \left[\log \left(\frac{p(x_1:x_T)}{q(x_1:x_T | x_0)} \right) \right] \end{aligned}$$

$$q(x_1 : x_T | x_0) = q(x_1 | x_0) q(x_2 | x_1) \cdots q(x_T | x_{T-1})$$

进行 ELBO 分析：

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \\
&= \log \int \frac{p(\mathbf{x}_{0:T}) q(\mathbf{x}_{1:T} | \mathbf{x}_0)}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} d\mathbf{x}_{1:T} \\
&= \log \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\frac{p(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \right] \\
&\geq \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) \prod_{t=1}^T p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{\prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1) \prod_{t=2}^T p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_T | \mathbf{x}_{T-1}) \prod_{t=1}^{T-1} q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1) \prod_{t=1}^{T-1} p_{\theta}(\mathbf{x}_t | \mathbf{x}_{t+1})}{q(\mathbf{x}_T | \mathbf{x}_{T-1}) \prod_{t=1}^{T-1} q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_T | \mathbf{x}_{T-1})} \right] + \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_{\theta}(\mathbf{x}_t | \mathbf{x}_{t+1})}{q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} [\log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)] + \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T | \mathbf{x}_{T-1})} \right] + \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_{\theta}(\mathbf{x}_t | \mathbf{x}_{t+1})}{q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} [\log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)] + \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T | \mathbf{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p_{\theta}(\mathbf{x}_t | \mathbf{x}_{t+1})}{q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_1 | \mathbf{x}_0)} [\log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)] + \mathbb{E}_{q(\mathbf{x}_{T-1}, \mathbf{x}_T | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T | \mathbf{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1} | \mathbf{x}_0)} \left[\log \frac{p_{\theta}(\mathbf{x}_t | \mathbf{x}_{t+1})}{q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \\
&= \underbrace{\mathbb{E}_{q(\mathbf{x}_1 | \mathbf{x}_0)} [\log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\mathbf{x}_{T-1} | \mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{x}_{T-1}) \| p(\mathbf{x}_T))]}_{\text{prior matching term}} \\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\mathbf{x}_{t-1}, \mathbf{x}_{t+1} | \mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_t | \mathbf{x}_{t-1}) \| p_{\theta}(\mathbf{x}_t | \mathbf{x}_{t+1}))]}_{\text{consistency term}}
\end{aligned}$$

上述优化是 suboptimal 的，引入新的概率：

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t | \mathbf{x}_0)}{q(\mathbf{x}_{t-1} | \mathbf{x}_0)} \quad (46)$$

Armed with this new equation, we can retry the derivation resuming from the ELBO in Equation 37:

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \right] \quad (47)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) \prod_{t=1}^T p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{\prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \quad (48)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1) \prod_{t=2}^T p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_1 | \mathbf{x}_0) \prod_{t=2}^T q(\mathbf{x}_t | \mathbf{x}_{t-1})} \right] \quad (49)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1) \prod_{t=2}^T p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_1 | \mathbf{x}_0) \prod_{t=2}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0)} \right] \quad (50)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p_{\theta}(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_1 | \mathbf{x}_0)} + \log \prod_{t=2}^T \frac{p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0)} \right] \quad (51)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_1 | \mathbf{x}_0)} + \log \prod_{t=2}^T \frac{p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{\frac{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t | \mathbf{x}_0)}{q(\mathbf{x}_{t-1} | \mathbf{x}_0)}} \right] \quad (52)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_1 | \mathbf{x}_0)} + \log \prod_{t=2}^T \frac{p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{\frac{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t | \mathbf{x}_0)}{q(\mathbf{x}_{t-1} | \mathbf{x}_0)}} \right] \quad (53)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_1 | \mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1 | \mathbf{x}_0)}{q(\mathbf{x}_T | \mathbf{x}_0)} + \log \prod_{t=2}^T \frac{p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)} \right] \quad (54)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T) p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)}{q(\mathbf{x}_T | \mathbf{x}_0)} + \sum_{t=2}^T \log \frac{p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)} \right] \quad (55)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} [\log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)] + \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T | \mathbf{x}_0)} \right] + \sum_{t=2}^T \mathbb{E}_{q(\mathbf{x}_{1:T} | \mathbf{x}_0)} \left[\log \frac{p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)} \right] \quad (56)$$

$$= \mathbb{E}_{q(\mathbf{x}_1 | \mathbf{x}_0)} [\log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)] + \mathbb{E}_{q(\mathbf{x}_T | \mathbf{x}_0)} \left[\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T | \mathbf{x}_0)} \right] + \sum_{t=2}^T \mathbb{E}_{q(\mathbf{x}_t, \mathbf{x}_{t-1} | \mathbf{x}_0)} \left[\log \frac{p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)}{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)} \right] \quad (57)$$

$$= \underbrace{\mathbb{E}_{q(\mathbf{x}_1 | \mathbf{x}_0)} [\log p_{\theta}(\mathbf{x}_0 | \mathbf{x}_1)]}_{\text{reconstruction term}} - \underbrace{D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{x}_0) \| p(\mathbf{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(\mathbf{x}_t | \mathbf{x}_0)} [D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t))]}_{\text{denoising matching term}} \quad (58)$$

$$\begin{aligned} & \text{maxmise } \mathbb{E}_{q(x_1 | x_0)} [\log p_{\theta}(x_0 | x_1)] - KL[q(x_T | x_0) \| p(x_T)] \\ & \quad - \sum_{t=2}^T \mathbb{E}_{q(x_t | x_0)} [KL[q(x_{t-1} | x_t, x_0) \| p_{\theta}(x_{t-1} | x_t)]] \end{aligned}$$

注意到 $\Gamma_{PMT} = -KL[q(x_T | x_0) \| p(x_T)]$ 与参数 θ 无关 (Noise 和 Noise 的 KL)。因此：

$$\begin{aligned} & \text{maxmise } \mathbb{E}_{q(x_1 | x_0)} [\log p_{\theta}(x_0 | x_1)] - \sum_{t=2}^T \mathbb{E}_{q(x_t | x_0)} [KL[q(x_{t-1} | x_t, x_0) \| p_{\theta}(x_{t-1} | x_t)]] \\ & \quad + \Gamma_{PMT} \end{aligned}$$

$q(x_{t-1} | x_t, x_0)$ 此项可以进行改写：

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) q(\mathbf{x}_{t-1} | \mathbf{x}_0)}{q(\mathbf{x}_t | \mathbf{x}_0)} \quad (71)$$

$$= \frac{\mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t) \mathbf{I}) \mathcal{N}(\mathbf{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0, (1 - \bar{\alpha}_{t-1}) \mathbf{I})}{\mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I})} \quad (72)$$

$$\propto \exp \left\{ - \left[\frac{(\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}_{t-1})^2}{2(1 - \alpha_t)} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0)^2}{2(1 - \bar{\alpha}_{t-1})} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^2}{2(1 - \bar{\alpha}_t)} \right] \right\} \quad (73)$$

$$= \exp \left\{ - \frac{1}{2} \left[\frac{(\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}_{t-1})^2}{1 - \alpha_t} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0)^2}{1 - \bar{\alpha}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^2}{1 - \bar{\alpha}_t} \right] \right\} \quad (74)$$

$$= \exp \left\{ - \frac{1}{2} \left[\frac{(-2\sqrt{\alpha_t} \mathbf{x}_t \mathbf{x}_{t-1} + \alpha_t \mathbf{x}_{t-1}^2)}{1 - \alpha_t} + \frac{(\mathbf{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_{t-1} \mathbf{x}_0)}{1 - \bar{\alpha}_{t-1}} + C(\mathbf{x}_t, \mathbf{x}_0) \right] \right\} \quad (75)$$

$$\propto \exp \left\{ - \frac{1}{2} \left[- \frac{2\sqrt{\alpha_t} \mathbf{x}_t \mathbf{x}_{t-1}}{1 - \alpha_t} + \frac{\alpha_t \mathbf{x}_{t-1}^2}{1 - \alpha_t} + \frac{\mathbf{x}_{t-1}^2}{1 - \bar{\alpha}_{t-1}} - \frac{2\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_{t-1} \mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right] \right\} \quad (76)$$

$$= \exp \left\{ - \frac{1}{2} \left[\left(\frac{\alpha_t}{1 - \alpha_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t} \mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1} \right] \right\} \quad (77)$$

$$= \exp \left\{ - \frac{1}{2} \left[\frac{\alpha_t(1 - \bar{\alpha}_{t-1}) + 1 - \alpha_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t} \mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1} \right] \right\} \quad (78)$$

$$= \exp \left\{ - \frac{1}{2} \left[\frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t} \mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1} \right] \right\} \quad (79)$$

$$= \exp \left\{ - \frac{1}{2} \left[\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \mathbf{x}_{t-1}^2 - 2 \left(\frac{\sqrt{\alpha_t} \mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) \mathbf{x}_{t-1} \right] \right\} \quad (80)$$

$$= \exp \left\{ - \frac{1}{2} \left(\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \right) \left[\mathbf{x}_{t-1}^2 - 2 \frac{\left(\frac{\sqrt{\alpha_t} \mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right)}{\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}} \mathbf{x}_{t-1} \right] \right\} \quad (81)$$

$$= \exp \left\{ - \frac{1}{2} \left(\frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \right) \left[\mathbf{x}_{t-1}^2 - 2 \frac{\left(\frac{\sqrt{\alpha_t} \mathbf{x}_t}{1 - \alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0}{1 - \bar{\alpha}_{t-1}} \right) (1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_{t-1} \right] \right\} \quad (82)$$

$$= \exp \left\{ - \frac{1}{2} \left(\frac{1}{\frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}} \right) \left[\mathbf{x}_{t-1}^2 - 2 \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1}) \mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t) \mathbf{x}_0}{1 - \bar{\alpha}_t} \mathbf{x}_{t-1} \right] \right\} \quad (83)$$

$$\propto \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1}) \mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t) \mathbf{x}_0}{1 - \bar{\alpha}_t}}_{\mu_q(\mathbf{x}_t, \mathbf{x}_0)}, \underbrace{\frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{I}}_{\Sigma_q(t)}) \quad (84)$$

$$q(x_{t-1} | x_t, x_0) = \mathcal{N} \left(x_{t-1}; \underbrace{\frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t x_0 + \sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1}) x_t}{1 - \bar{\alpha}_t}}_{\mu_q(x_t, x_0)}, \underbrace{\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t \mathbf{I}}_{\Sigma_q(t)} \right)$$

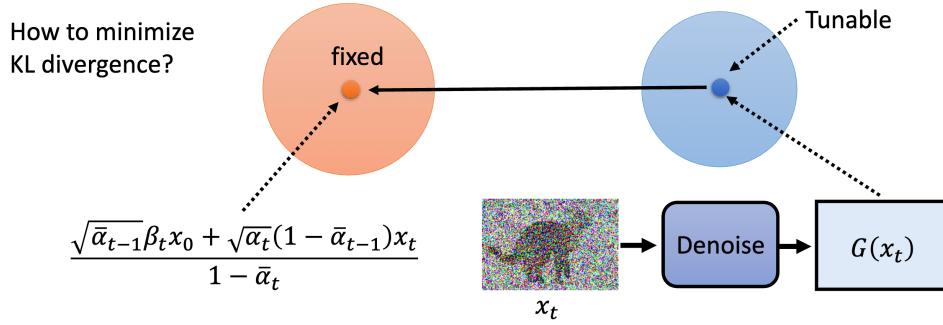
考慮 KL 的解析解：

$$\begin{aligned} &KL[\mathcal{N}(x; \mu_x, \Sigma_x) \| \mathcal{N}(y; \mu_y, \Sigma_y)] \\ &= \frac{1}{2} \left[\log \frac{|\Sigma_y|}{|\Sigma_x|} - d + \text{tr}[\Sigma_y^{-1} \Sigma_x] + (\mu_y - \mu_x)^\top \Sigma_y^{-1} (\mu_y - \mu_x) \right] \end{aligned}$$

$$KL[q(x_{t-1} | x_t, x_0) \| p_\theta(x_{t-1} | x_t)]$$

$$\begin{aligned}
 & \arg \min_{\theta} D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)) \\
 &= \arg \min_{\theta} D_{\text{KL}}(\mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)) \| \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}, \boldsymbol{\Sigma}_q(t))) \\
 &= \arg \min_{\theta} \frac{1}{2} \left[\log \frac{|\boldsymbol{\Sigma}_q(t)|}{|\boldsymbol{\Sigma}_{\theta}(t)|} - d + \text{tr}(\boldsymbol{\Sigma}_q(t)^{-1} \boldsymbol{\Sigma}_{\theta}(t)) + (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q) \right] \\
 &= \arg \min_{\theta} \frac{1}{2} [\log 1 - d + d + (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)] \\
 &= \arg \min_{\theta} \frac{1}{2} [(\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)^T \boldsymbol{\Sigma}_q(t)^{-1} (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)] \\
 &= \arg \min_{\theta} \frac{1}{2} [(\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)^T (\sigma_q^2(t) \mathbf{I})^{-1} (\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q)] \\
 &= \arg \min_{\theta} \frac{1}{2 \sigma_q^2(t)} [\|\boldsymbol{\mu}_{\theta} - \boldsymbol{\mu}_q\|_2^2]
 \end{aligned}$$

$$-\sum_{t=2}^T \mathbb{E}_{q(x_t | x_0)} [KL(q(x_{t-1} | x_t, x_0) || P(x_{t-1} | x_t))]$$



已知

$$\mu_q = \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t x_0 + \sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1}) x_t}{1 - \bar{\alpha}_t}$$

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \varepsilon \quad \Rightarrow \quad x_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \varepsilon}{\sqrt{\bar{\alpha}_t}}$$

$$\begin{aligned}
 \mu_q &= \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \varepsilon}{\sqrt{\bar{\alpha}_t}} + \sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1}) x_t}{1 - \bar{\alpha}_t} \\
 &= \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon \right)
 \end{aligned}$$

只有 ε 不知，因此 $\epsilon_{\theta}(x_t, t) = \varepsilon$

Algorithm 1 Training

```

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

```

Algorithm 2 Sampling

```

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

```

最后这里的 $\sigma_t z$ 可能是类似于 NLP 中的 Sample etc. DDPM 本身是一个 Autoregressive 的模型，即可以一步到位。但是可能一步到位的效果不好，因此去选择分布采样。这就是为什么要加入 $\sigma_t z$ 。

