扩散模型奠基与DDPM

前向过程

1. 数据分布：$q(x^{(0)})$
2. 目标分布：一个容易解析的简单分布$\pi\left(x^{(T)}\right)$
3. 马尔可夫内核T_π：一个一般性的马尔可夫矩阵，将当前状态映射到下一个状态的概率

$$q\left(x^{(t)}|x^{(t-1)}\right)=T_\pi\left(x^{(t)}|x^{(t-1)};\beta_t\right)$$
$$q\left(x^{(0\cdots T)}\right)=q\left(x^{(0)}\right)\prod_{t=1}^{T}q\left(x^{(t)}|x^{(t-1)}\right)$$

反向过程

1. 原始分布：简单分布$\pi\left(x^{(T)}\right)$
2. 目标分布：数据分布

$$p\left(x^{(T)}\right)=\pi\left(x^{(T)}\right)$$
$$p\left(x^{(0\cdots T)}\right)=p\left(x^{(T)}\right)\prod_{t=1}^{T}p\left(x^{(t-1)}|x^{(t)}\right)$$

关键假设：当每一次加噪的幅度非常小，如果正向过程是高斯分布（或者二项分布），则反向过程也是高斯分布（或者二项分布）因此我们可以直接根据非常少的参数就能够把反向的过程确定下来。

详解反向过程

$$p\left(x^{(0\cdots T)}\right)=p\left(x^{(T)}\right)\prod_{t=1}^{T}p\left(x^{(t-1)}|x^{(t)}\right)$$

$$\begin{aligned}
p\left(x^{(0)}\right)&=\int dx^{(1)}\int dx^{(2)}\cdots\int x^{(T)}p\left(x^{(0\cdots T)}\right)\\
&=\int dx^{(1\cdots T)}p\left(x^{(0\cdots T)}\right)\frac{q\left(x^{(1\cdots T)}|x^{(0)}\right)}{q\left(x^{(1\cdots T)}|x^{(0)}\right)}\\
&=\int dx^{(1\cdots T)}q\left(x^{(1\cdots T)}|x^{(0)}\right)\frac{p\left(x^{(0\cdots T)}\right)}{q\left(x^{(1\cdots T)}|x^{(0)}\right)}\\
&=\int dx^{(1\cdots T)}q\left(x^{(1\cdots T)}|x^{(0)}\right)\cdot p\left(x^{(T)}\right)\prod_{t=1}^{T}\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)}
\end{aligned}$$

对数似然及其ELBO

现在我们已经有了$p\left(x^{(0)}\right)$的一个比较好的解析方法，因此可以用来比较模型的输出值$p\left(x^{(0)}\right)$和真实数据$q\left(x^{(0)}\right)$之间的匹配程度，用对数似然函数即可：

$$
\begin{aligned}
L &= \int dx^{(0)} q\left(x^{(0)}\right) \log p\left(x^{(0)}\right) \\
&= \int dx^{(0)} q\left(x^{(0)}\right) \log \left[\int dx^{(1\cdots T)} q\left(x^{(1\cdots T)} | x^{(0)}\right) \cdot p\left(x^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right]
\end{aligned}
$$

由Jenson不等式可知，对于任意的上凸函数，都有$f[E(x)] \geq E[f(x)]$，而log函数就是一个上凸函数，因此我们可以得到对数似然函数的下界：

$$
\begin{aligned}
L &= \int dx^{(0)} q\left(x^{(0)}\right) \log E_{x^{(1\cdots T)} \sim q\left(x^{(1\cdots T)} | x^{(0)}\right)} \left[p\left(x^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] \\
&\geq \int dx^{(0)} q\left(x^{(0)}\right) E_{x^{(1\cdots T)} \sim q\left(x^{(1\cdots T)} | x^{(0)}\right)} \log \left[p\left(x^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] \\
&= \int dx^{(0)} dx^{(1\cdots T)} q\left(x^{(0)}\right) q\left(x^{(1\cdots T)} | x^{(0)}\right) \log \left[p\left(x^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] \\
&= \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[p\left(x^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] \\
&= \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \left[\sum_{t=1}^{T} \log \frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)} + \log p\left(x^{(T)}\right)\right] \\
&= K
\end{aligned}
$$

接下来我们可以对这个下界K继续进行处理：由(1)式可知，我们可以把$q\left(x^{(0\cdots T)}\right)$转换为$q\left(x^{(0)}\right) \prod_{t=1}^{T} q\left(x^{(t)} | x^{(t-1)}\right)$，后面的条件概率刚好可以和前面的积分抵消形成边缘概率分布，最终得到：

$$
\begin{aligned}
K &= \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \sum_{t=1}^{T} \log \left[\frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] + \int dx^{(T)} q\left(x^{(T)}\right) \log p\left(x^{(T)}\right) \\
&= \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \sum_{t=1}^{T} \log \left[\frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] + \int dx^{(T)} \pi\left(x^{(T)}\right) \log \pi\left(x^{(T)}\right) \\
&= \sum_{t=1}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] - H_p\left(X^{(T)}\right)
\end{aligned}
$$

接下来我们要考虑边界点，这需要进行单独考虑，就好比抓住绳子的两端之后要并在一起才能把绳子完美地对折：

$$
p\left(x^{(0)} | x^{(1)}\right) = q\left(x^{(1)} | x^{(0)}\right) \frac{\pi\left(x^{(0)}\right)}{\pi\left(x^{(1)}\right)} = T_\pi\left(x^{(0)} | x^{(1)}; \beta_1\right)
$$

$$
\begin{aligned}
K &= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] + \int dx^{(0)} \int dx^{(1)} q\left(x^{(0)}, x^{(1)}\right) \log \left[\frac{p\left(x^{(0)} | x^{(1)}\right)}{q\left(x^{(1)} | x^{(0)}\right)}\right] - H_p\left(X^{(T)}\right) \\
&= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] + \int dx^{(0)} \int dx^{(1)} q\left(x^{(0)}, x^{(1)}\right) \log \left[\frac{q\left(x^{(1)} | x^{(0)}\right) \pi\left(x^{(0)}\right)}{q\left(x^{(1)} | x^{(0)}\right) \pi\left(x^{(1)}\right)}\right] - H_p\left(X^{(T)}\right) \\
&= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)} | x^{(t)}\right)}{q\left(x^{(t)} | x^{(t-1)}\right)}\right] + \int dx^{(0)} \int dx^{(1)} q\left(x^{(0)}, x^{(1)}\right) \log \left[\frac{\pi\left(x^{(0)}\right)}{\pi\left(x^{(1)}\right)}\right] - H_p\left(X^{(T)}\right)
\end{aligned}
$$

由于前向加噪的初始步长非常小，实际上通过积分得到了两个Entropy可以直接消去，最终得到：

$$
\begin{aligned}
K &= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}\right)}\right] - H_p\left(X^{(T)}\right) \\
&= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t)}|x^{(t-1)}, x^{(0)}\right)}\right] - H_p\left(X^{(T)}\right) \\
&= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)} \frac{q\left(x^{(t-1)}|x^{(0)}\right)}{q\left(x^{(t)}|x^{(0)}\right)}\right] - H_p\left(X^{(T)}\right) \\
&= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)}\right] + \sum_{t=2}^{T}\left[H_q\left(X^{(t)}|X^{(0)}\right) - H_q\left(X^{(t-1)}|X^{(0)}\right)\right] - H_p\left(X^{(T)}\right) \\
&= \sum_{t=2}^{T} \int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)}\right] + H_q\left(X^{(T)}|X^{(0)}\right) - H_q\left(X^{(1)}|X^{(0)}\right) - H_p\left(X^{(T)}\right)
\end{aligned}
$$

接下来处理前面的大块头：由于log函数里面只和$x^{(t)}, x^{(t-1)}, x^{(0)}$有关，所以对于每一项积分，我们都可以把其他的积分全部消掉变成1:

$$
\begin{aligned}
&\int dx^{(0\cdots T)} q\left(x^{(0\cdots T)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)}\right] \\
&= \int dx^{(0)} dx^{(t-1)} dx^{(t)} q\left(x^{(0)}\right) q\left(x^{(t-1)}\right) q\left(x^{(t)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)}\right] \\
&= \int dx^{(0)} dx^{(t-1)} dx^{(t)} q\left(x^{(0)}, x^{(t)}\right) q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)}\right] \\
&= \int dx^{(0)} dx^{(t)} q\left(x^{(0)}, x^{(t)}\right) \int dx^{(t-1)} q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) \log \left[\frac{p\left(x^{(t-1)}|x^{(t)}\right)}{q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right)}\right] \\
&= -\int dx^{(0)} dx^{(t)} q\left(x^{(0)}, x^{(t)}\right) D_{KL}\left(q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) \parallel p\left(x^{(t-1)}|x^{(t)}\right)\right)
\end{aligned}
$$

因此我们可以用KL散度来表示K：

$$
\begin{aligned}
K = &-\sum_{t=2}^{T} \int dx^{(0)} dx^{(t)} q\left(x^{(0)}, x^{(t)}\right) D_{KL}\left(q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) \parallel p\left(x^{(t-1)}|x^{(t)}\right)\right) \\
&+ H_q\left(X^{(T)}|X^{(0)}\right) - H_q\left(X^{(1)}|X^{(0)}\right) - H_p\left(X^{(T)}\right)
\end{aligned}
$$

由于后面几个熵的值都是常数，我们只能尝试通过最小化$D_{KL}\left(q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) \parallel p\left(x^{(t-1)}|x^{(t)}\right)\right)$使得$K$尽可能达到最大，因此$L$的下界尽可能增大。

前向扩散核、反向扩散核

在前向扩散的时候，我们可以自定义加噪公式：

$$
q\left(x^{(t)}|x^{(t-1)}\right) = \mathcal{N}\left(x^{(t)}; \sqrt{1-\beta_t} \cdot x^{(t-1)}, \mathbf{I}\beta_t\right)
$$

$$
x^{(t)} = \sqrt{1-\beta_t} \cdot x^{(t-1)} + \beta_t \cdot \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0, \mathbf{I})
$$

这里的β_t是自定义的，你可以选择线性增长或者三角函数型增长。通过每一步加噪，我们将分布的均值逐渐往0逼近，并且方差越来越大，最后形成一个$\mathcal{N}\left(x^{(T)}; 0, \mathbf{I}\right)$的标准正态分布。

反向过程中需要学习两个参数：

$$
p\left(x^{(t-1)}|x^{(t)}\right) = \mathcal{N}\left(x^{(t-1)}; f_\mu\left(x^{(t)}; t\right), f_\Sigma\left(x^{(t)}; t\right)\right)
$$

即每一次反向过程中的均值和方差。

简化表达 (DDPM开始发力)

1. α的引入

很显然，如果我们真的按照每一步加噪、去噪来训练模型的话，那么模型训练的时间相当长，因为每一次我都要以上一次加噪（去噪）的结果为输入，因此在前面的时间步进行训练的时候，后面的时间步只能等待。

我们发现从$x^{(0)}$到$x^{(t)}$的累积过程可以递归展开：

$$x^{(1)} = \sqrt{1-\beta_1} \cdot x^{(0)} + \beta_1 \cdot \epsilon_1$$

$$\begin{aligned} x^{(2)} &= \sqrt{1-\beta_2} \cdot x^{(1)} + \beta_2 \cdot \epsilon_2 \\ &= \sqrt{1-\beta_2} \cdot \left(\sqrt{1-\beta_1} \cdot x^{(0)} + \beta_1 \cdot \epsilon_1\right) + \beta_2 \cdot \epsilon_2 \\ &= \sqrt{(1-\beta_2)(1-\beta_1)} \cdot x^{(0)} + \sqrt{(1-\beta_2)\beta_1} \cdot \epsilon_1 + \sqrt{\beta_2} \cdot \epsilon_2 \end{aligned}$$

由于ϵ_1, ϵ_2相互独立并且服从$\mathcal{N}(0, \mathbf{I})$，因此我们可以把他们俩合并成一个高斯分布：

$$\begin{aligned} \left(\sqrt{(1-\beta_2)\beta_1} \cdot \epsilon_1 + \sqrt{\beta_2} \cdot \epsilon_2\right) &\sim \mathcal{N}\left(0, [(1-\beta_2)\beta_1 + \beta_2]\mathbf{I}\right) \\ &= \mathcal{N}(0, [1-(1-\beta_1)(1-\beta_2)]\mathbf{I}) \end{aligned}$$

接下来就可以推广到t步：定义$\alpha_t = 1 - \beta_t$，$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$，则$x^{(t)}$可以表示为：

$$x^{(t)} = \sqrt{\bar{\alpha}_t} \cdot x^{(0)} + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon,\ \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

$$q\left(x^{(t)} | x^{(0)}\right) = \mathcal{N}\left(x^{(t)}; \sqrt{\bar{\alpha}_t} \cdot x^{(0)}, (1-\bar{\alpha}_t)\mathbf{I}\right)$$

这个表达式仅依赖于$\bar{\alpha}_t$，即α_t的累积乘积，避免了逐步计算噪声项的复杂性。

2. D_{KL}在高斯分布条件下的简化与损失函数

上文已经给出，我们的目标是**最小化$D_{KL}\left(q\left(x^{(t-1)} | x^{(t)}, x^{(0)}\right) \parallel p_\theta\left(x^{(t-1)} | x^{(t)}\right)\right)$使得$K$尽可能达到最大**，其中$\theta$表示神经网络的参数

对于两个高斯分布$\mathcal{N}(\mu_1; \sigma^2\mathbf{I}), \mathcal{N}(\mu_2; \sigma^2\mathbf{I})$，他们之间的$D_{KL}$可以用以下方式表示：

$$D_{KL}\left(\mathcal{N}(\mu_1; \sigma^2\mathbf{I}) \parallel \mathcal{N}(\mu_2; \sigma^2\mathbf{I})\right) = \frac{1}{2\sigma^2}\|\mu_1 - \mu_2\|^2$$

接下来我们需要计算前向过程的真实后验分布，即$q\left(x^{(t-1)} | x^{(t)}, x^{(0)}\right)$，它也遵循高斯分布，不过比较复杂：

$$q\left(x^{(t-1)} | x^{(t)}, x^{(0)}\right) = \mathcal{N}\left(x^{(t-1)}; \tilde{\mu}_t\left(x^{(t)}, x^{(0)}\right), \tilde{\beta}_t\mathbf{I}\right)$$

在这里，均值和方差都有解析解，接下来推导其过程：

目标是求$q\left(x^{(t-1)} | x^{(t)}, x^{(0)}\right)$，即已知$x^{(t)}$和$x^{(0)}$时，$x^{(t-1)}$的条件分布。

根据贝叶斯定义：

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

由于前向过程是马尔可夫链，$q\left(x^{(t-1)} | x^{(t)}, x^{(0)}\right) = q\left(x^{(t-1)} | x^{(t)}\right)$，因此：

$$q\left(x^{(t-1)}|x^{(t)},x^{(0)}\right)\propto q\left(x^{(t)}|x^{(t-1)}\right)\cdot q\left(x^{(t-1)}|x^{(0)}\right)$$

接下来把各个高斯分布显式地写出：

1. *前向单步转移概率：*

$$q\left(x^{(t)}|x^{(t-1)}\right)\propto\exp\left(-\frac{\left\|x^{(t)}-\sqrt{\alpha_t}x^{(t-1)}\right\|^2}{2\beta_t}\right)$$

2. *从$x^{(0)}$到$x^{(t-1)}$的累积分布：*

$$q\left(x^{(t-1)}|x^{(0)}\right)\propto\exp\left(-\frac{\left\|x^{(t-1)}-\sqrt{\bar{\alpha}_{t-1}}x^{(0)}\right\|^2}{2(1-\bar{\alpha}_{t-1})}\right)$$

3. *联合分布的指数项，即后验分布的指数项为：*

$$-\frac{\left\|x^{(t)}-\sqrt{\alpha_t}x^{(t-1)}\right\|^2}{2\beta_t}-\frac{\left\|x^{(t-1)}-\sqrt{\bar{\alpha}_{t-1}}x^{(0)}\right\|^2}{2(1-\bar{\alpha}_{t-1})}$$

将指数项展开并合并同类项之后可以得到关于$x^{(t-1)}$的二次项系数和一次项系数：

1. *二次项系数：*

$$-\frac{1}{2}\left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\right)\left(x^{(t-1)}\right)^2$$

2. *一次项系数：*

$$-\frac{1}{2}\left(\frac{\sqrt{\alpha_t}}{\beta_t}x^{(t)}+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x^{(0)}\right)x^{(t-1)}$$

那么分别对应到一个典型高斯分布的指数项形式：

$$-\frac{(x-\mu)^2}{2\sigma^2}=-\frac{x^2}{2\sigma^2}+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}$$

通过对比，就可以得到后验分布的均值和方差：

1. *方差：*

$$\frac{1}{\tilde{\beta}_t}=\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\Longrightarrow\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

2. *均值：*

$$\frac{\tilde{\mu}_t}{\tilde{\beta}_t}=\frac{\sqrt{\alpha_t}}{\beta_t}x^{(t)}+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x^{(0)}\Longrightarrow\tilde{\mu}_t=\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x^{(t)}+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x^{(0)}$$

再根据前向过程$x^{(t)}=\sqrt{\bar{\alpha}_t}\cdot x^{(0)}+\sqrt{1-\bar{\alpha}_t}\cdot\epsilon$，解得：

$$x^{(0)}=\frac{1}{\sqrt{\bar{\alpha}_t}}\left(x^{(t)}-\sqrt{1-\bar{\alpha}_t}\cdot\epsilon\right)$$

将$x^{(0)}$带入均值表达式可得：

$$\tilde{\mu}_t=\frac{1}{\sqrt{\alpha_t}}\left(x^{(t)}-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right)$$

因此，如果神经网络预测的反向过程的均值μ_θ为：

$$\mu_\theta\left(x^{(t)}, t\right) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x^{(t)} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x^{(t)}, t\right)\right)$$

则D_{KL}就可以简化为：

$$\begin{aligned}
&D_{KL}\left(q\left(x^{(t-1)}|x^{(t)}, x^{(0)}\right) \parallel p_\theta\left(x^{(t-1)}|x^{(t)}\right)\right) \\
=&\frac{1}{2\sigma^2}\left\|\tilde{\mu}_t - \mu_\theta\left(x^{(t)}, t\right)\right\|^2 \\
=&\frac{\beta_t^2}{2\tilde{\beta}_t\alpha_t(1-\bar{\alpha}_t)}\left\|\epsilon - \epsilon_\theta\left(x^{(t)}, t\right)\right\|^2
\end{aligned}$$

忽略权重系数之后，**最终损失函数**为：

$$\mathcal{L}_t(\theta) = \mathbf{E}_{x^{(0)},\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(x^{(t)}, t\right)\right\|^2\right]$$

1. **高斯假设**：前向和反向过程均为高斯分布，且方差固定。
2. **均值匹配**：KL散度简化为均值差的MSE。
3. **噪声预测**：通过参数化技巧，将均值预测转换为对噪声ϵ的直接预测。
4. **损失函数**：最终损失函数为预测噪声与真实噪声的MSE。