Denoising Diffusion Probabilistic Models

xinli*

September 15, 2024

Abstract

在学习 DDPM 和扩散模型的过程中，我记录了一些思考和笔记，以加深对这些概念的理解，并方便将来在遗忘时能更好地查阅和复习相关信息。（仅供参考）

1 Preliminary

在开始学习 DDPM 之前，我们先学习一些预备知识，方便我们接下来更好的理解。

1.1 正态分布的表示方法

$$N(x; \mu; \sigma^2) \tag{1}$$

公式(1)表示随机变量 x 服从一个均值为 μ，方差为 σ^2 的分布。具体来说，这意味着：**μ 是正态分布的均值**，表示数据分布的中心位置和期望值。**σ^2 是正态分布的方差**，表示数据的波动性。标准差 σ 是方差的平方根。它的概率密度函数通常可以表示为：

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2}$$

1.2 重参数化技巧

如果你了解变分自编码器 VAE 的话，你应该对重参数化不陌生。**重参数化技巧的核心思想是将不可导的采样过程转化为可导的操作。**具体来说，假设我们有一个随机变量 z，其分布参数化为均值 μ 和方差 σ^2，即：

$$z \sim \mathcal{N}\left(\mu, \sigma^2\right) \tag{3}$$

直接从这个分布中采样的过程不可导，但我们可以通过以下方法重参数化，使其可导：

1. 首先，从标准正态分布中采样一个变量 ϵ:

$$\epsilon \sim \mathcal{N}(0, 1) \tag{4}$$

2. 然后，通过一个线性变换，将 ϵ 转换为具有均值 μ 和方差 σ^2 的目标分布：

$$z = \mu + \sigma \cdot \epsilon \tag{5}$$

*https://github.com/xinli2008

这里，μ 和 σ 是模型参数，它们的梯度是可导的，因为 ϵ 是从固定的标准正态分布中采样的，与模型参数无关。

1.3 条件概率分布和联合概率分布

联合概率分布表示两个或多个随机变量随机发生的概率。对于两个随机变量 X 和 Y，其联合概率分布 $P(X,Y)$ 表示同时观察到 X 和 Y 取某些值的概率。**而条件概率表示在已知一个事件发生的前提下，另一个事件发生的概率**。对于随机变量 Y 在已知 X 的情况下的条件概率分布表示为 $P(Y \mid X)$。

$$P(A,B,C) = P(C \mid B,A) \cdot P(B \mid A) \cdot P(A) \tag{6}$$

这个公式表示事件 A、B、C 同时发生的概率。我们可以一步一步来思考，先计算事件 A 发生的概率 $P(A)$，随后在计算事件 A 发生后，事件 B 发生的概率 $P(B \mid A)$, 最后在计算 A 和 B 发生后，事件 C 发生的条件概率 $P(C \mid B,A)$。

$$P(B,C \mid A) = P(B \mid A)P(C \mid A,B) \tag{7}$$

这个公式表示在给定事件 A 发生的情况下，事件 B 和 C 同时发生的概率。我们同样可以将它分解成两个部分：先计算 A 发生的概率下，事件 B 发生的概率 $P(B \mid A)$, 在计算 A 和 B 都发生的情况下，事件 C 发生的概率 $P(C \mid A,B)$。

1.4 KL 散度的介绍

KL 散度（Kullback-Leibler Divergence）是用来衡量两个概率分布 $p(x)$ 和 $q(x)$ 之间的差异。对于离散分布，KL 散度定义为：

$$KL(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{8}$$

对于连续分布，这个定义对应的公式是：

$$KL(p \parallel q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{9}$$

这个公式衡量的是分布 p 相对于分布 q 的“信息损失”。

现在，我们考虑两个单变量高斯分布 $p(x)$ 和 $q(x)$，它们的概率密度函数分别是：

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \tag{10}$$

$$q(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right) \tag{11}$$

我们要计算的是这两个分布之间的 KL 散度：

$$KL(p \parallel q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{12}$$

将上面 $p(x)$ 和 $q(x)$ 的表达式代入 KL 散度的公式中：

$$KL(p \parallel q) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \log \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)} dx \tag{13}$$

首先化简对数表达式：

$$\log \frac{\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)} = \log\left(\frac{\sqrt{2\pi\sigma_2^2}}{\sqrt{2\pi\sigma_1^2}} \cdot \exp\left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\right) \tag{14}$$

由于对数的性质，这个表达式可以进一步分解为两部分：

$$\log \frac{\sigma_2}{\sigma_1} + \left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \tag{15}$$

将化简后的对数表达式代入 KL 散度公式中：

$$KL(p \parallel q) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \left[\log \frac{\sigma_2}{\sigma_1} + \left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\right] dx \tag{16}$$

这个积分可以拆分为两部分：

$$\begin{aligned} KL(p \parallel q) = &\log \frac{\sigma_2}{\sigma_1} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) dx \\ &+ \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \left(\frac{(x-\mu_2)^2}{2\sigma_2^2} - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right) dx \end{aligned} \tag{17}$$

第一个积分是一个常数项的积分。**这个积分计算的是整个正态分布曲线下的面积，因为正态分布是标准化的，所以这个积分的结果是 1**：

$$\log \frac{\sigma_2}{\sigma_1} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) dx = \log \frac{\sigma_2}{\sigma_1} \cdot 1 = \log \frac{\sigma_2}{\sigma_1} \tag{18}$$

第二个积分更复杂，但通过高斯分布的性质可以简化。由于分布 $p(x)$ 是标准化的高斯分布，所以第二个积分的结果可以计算为：

$$\frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{19}$$

将两部分的结果相加，得到最终的 KL 散度公式：

$$KL(p \parallel q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{20}$$

2 Introduction

2.1 扩散模型的前向过程

给定一个从真实数据分布 $q(x)$ 中采样的数据点 x_0，我们定义一个前向扩散过程，其中我们在 T 步中向样本中加入少量的高斯噪声，生成一系列带噪声的样本 $x_1, ..., x_T$。数据样本 x_0 随着步数 t 的增大逐渐失去其可辨识度。当 $T \to \infty$ 时，X_T 等价于一个各向同性的高斯分布，也就是一个纯噪声图片。(**从这里我们可以知道，扩散模型的前向过程是一个加噪的过程，并且每一步加的噪声都是符合高斯分布的，当加到时间步 T 时，图像接近一个标准的高斯分布（纯噪声图片，标准正态分布）**：

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{21}$$

观察公式(21)，我们可以使用重参数化技巧(1.2)来在任意时间步长中进行采样。

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_{t-1} \tag{22}$$

为了接下来方便表示，我们令 $\alpha_t = 1 - \beta_t$。从(22)中，我们得到：

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \tag{23}$$

$$\mathbf{x}_t = \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \tag{24}$$

展开后，我们可以得到：

$$\mathbf{x}_t = \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{\alpha_t(1-\alpha_{t-1})}\boldsymbol{\epsilon}_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \tag{25}$$

注意这里，正态分布有个性质：**两个正态分布 $X \sim \mathcal{N}(\mu_1, \sigma_1)$ 和 $Y \sim \mathcal{N}(\mu_2, \sigma_2)$，叠加后的分布 $aX + bY$ 的均值为 $a\mu_1 + b\mu_2$，方差为 $a^2\sigma_1^2 + b^2\sigma_2^2$**。而 ϵ_{t-1} 和 ϵ_{t-2} 都是从均值为 0，方差为 1 的标准正态分布中进行采样的，所以 $\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\epsilon_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-1}$ 的均值为 0，方差为 $\alpha_t - \alpha_t\alpha_{t-1} + 1 - \alpha_t$，即 $1 - \alpha_t\alpha_{t-1}$，我们将后面两项给重建为 $\sqrt{1-\alpha_t\alpha_{t-1}}\epsilon_{t-2}$。

$$x_t = \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\tilde{\boldsymbol{\epsilon}}_{t-2} \tag{26}$$

为了接下来表示方便，我们将 $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$，我们将公式(26)依次递推，我们可以得到：

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{27}$$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{28}$$

因此，在扩散模型的前向过程中，给定 x_0 和时间步 t，可以直接计算出 x_t 而无需逐步模拟整个过程。在该步骤中，β **被设置成了不可学习的参数，范围在 [1e-4, 0.02] 之间，随着时间步 t 线性增加**，这也极大的简化了训练时候的优化目标。在代码中，T 一般需要设置为 1000，随着 t 的增加，样本 x_t 逐渐趋于标准正态分布（即纯噪声）。

2.2 扩散模型的逆向过程

如果我们可以逆转上述过程并从 $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ 中采样，那么我们将能够从高斯噪声输入中直接重建出真实样本。不幸的是，**我们无法轻易估计 $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$，因为它需要使用整个数据集，因此我们需要学习一个模型 p_θ 来不断的近似这个条件概率**(这里和 VAE 很类似，直接计算不出来，那就用神经网络拟合一个)，以便运行逆向扩散过程。

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \sigma_\theta\left(\mathbf{x}_t, t\right)\right) \tag{29}$$

值得注意的是，当**有条件 x_0 时，反向条件概率 $q(x_{t+1} \mid x_t, x_0)$ 是可处理的**:

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}\left(\mathbf{x}_t, \mathbf{x}_0\right), \tilde{\beta}_t\mathbf{I}\right) \tag{30}$$

$$q(x_{t-1} \mid x_t, x_0) = \frac{q(x_{t-1}, x_t, x_0)}{q(x_t \mid x_0) \cdot q(x_0)} \tag{31}$$

$$= \frac{q(x_t \mid x_{t-1}, x_0) \cdot q(x_{t-1} \mid x_0) \cdot q(x_0)}{q(x_t \mid x_0) \cdot q(x_0)} \tag{32}$$

$$= q(x_t \mid x_{t-1}, x_0)\frac{q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} \tag{33}$$

又因为分布 $q(\cdot)$ 是服从马尔科夫链，所以 $q(x_t \mid x_{t-1}, x_0) = q(x_t \mid x_{t-1})$，参考公式(2)和(28)，我们得到：

$$\propto \exp\left(-\frac{1}{2}\left(\frac{(x_t-\sqrt{\alpha_t}x_{t-1})^2}{\beta_t}+\frac{(x_{t-1}-\sqrt{\bar{\alpha}_{t-1}}x_0)^2}{1-\bar{\alpha}_{t-1}}-\frac{(x_t-\sqrt{\bar{\alpha}_t}x_0)^2}{1-\bar{\alpha}_t}\right)\right) \tag{34}$$

将平方拆开，我们得到：

$$= \exp\left(-\frac{1}{2}\left(\frac{x_t^2-2\sqrt{\alpha_t}x_t x_{t-1}+\alpha_t x_{t-1}^2}{\beta_t}+\frac{x_{t-1}^2-2\sqrt{\bar{\alpha}_{t-1}}x_{t-1}x_0+\bar{\alpha}_{t-1}x_0^2}{1-\bar{\alpha}_{t-1}}-\frac{x_t^2-2\sqrt{\bar{\alpha}_t}x_t x_0+\bar{\alpha}_t x_0^2}{1-\bar{\alpha}_t}\right)\right) \tag{35}$$

合并同类项，我们得到：

$$= \exp\left(-\frac{1}{2}\left(\left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\right)x_{t-1}^2-\left(\frac{2\sqrt{\alpha_t}}{\beta_t}x_t+\frac{2\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}x_0\right)x_{t-1}+C\left(X_0,X_t\right)\right)\right) \tag{36}$$

我们知道，概率密度函数 $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ 对应的分布的均值是 μ，方差是 σ^2。通过将(36)变换形式，我们可以得到它的均值是 $\frac{b}{-2a}$，方差为 $\frac{1}{a}$:

$$\tilde{\beta}_t = \frac{1}{\left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\right)} = \frac{1}{\left(\frac{\alpha_t-\bar{\alpha}_t+\beta_t}{\beta_t(1-\bar{\alpha}_{t-1})}\right)} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t \tag{37}$$

$$\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0) = \frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0 \Big/ \left(\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\right) \tag{38}$$

$$= \left(\frac{\sqrt{\alpha_t}}{\beta_t}\mathbf{x}_t+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}\mathbf{x}_0\right)\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t \tag{39}$$

$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 \tag{40}$$

根据公式(27), 我们可以推导出：

$$\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\epsilon_t\right) \tag{41}$$

将它带入到公式(40)，我们可以化简：

$$\tilde{\mu}_t = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t+\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\cdot\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t\right) \tag{42}$$

$$= \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right) \tag{43}$$

所以，**分布 $q\left(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{x}_0\right)$ 的均值 $\tilde{\mu}(x_t,x_0)$ 为 $\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\right)$，方差 $\tilde{\beta}_t$ 为 $\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\cdot\beta_t$**。我们可以观察到，表达式中不再含有 x_0，并且多了噪声项目，这位后面我们设计神经网络提供了基础。也就是说，在给定 x_0 的条件下，后验条件高斯分布 $q(x_{t-1}\mid x_t,x_0)$ 计算只与 x_t 和 z_t 有关。

2.3 目标数据分布的似然函数

扩散模型的训练目标（或者说是生成式模型的训练目标，其中包括 VAE）是在**给定观测数据的情况下，通过最大化似然来学习数据的分布，即让 p_θ 最大**。而扩散模型是通过最大化对数似然，来实现这一目标：

$$-\log p_\theta(\mathbf{x}_0) \leq -\log p_\theta(\mathbf{x}_0) + D_{\mathrm{KL}}(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\|p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)) \tag{44}$$

这里因为期望值的定义是对随机变量 X 在其分布 $q(x)$ 下加权求和或积分：$\mathbb{E}_{q(x)}[f(x)] = \int q(x)f(x)\,dx$。因此，我们可以结合 KL 散度的公式将上述公式化简为：

$$= -\log p_\theta(\mathbf{x}_0) + \mathbb{E}_{\mathbf{x}_{1:T}\sim q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{1:T})/p_\theta(\mathbf{x}_0)}\right] \tag{45}$$

$$= -\log p_\theta(\mathbf{x}_0) + \mathbb{E}_q\left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})} + \log p_\theta(\mathbf{x}_0)\right] \tag{46}$$

因为 $p_\theta(x_0)$ 显然与 q 无关，所以我们将它放在期望的外面：

$$-\log p_\theta(\mathbf{x}_0) = \mathbb{E}_q\left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] \tag{47}$$

$$\text{Let } \mathcal{L}_{\mathrm{VLB}} = \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] \geq -\mathbb{E}_{q(\mathbf{x}_0)}\log p_\theta(\mathbf{x}_0) \tag{48}$$

接下来，我们只需要最小化 L_{VLB}。L_{VLB} **最小，则** $-logp_\theta(x_0)$ **最小**。

$$L_{VLB} = \mathbb{E}_{q(x_{0:T})}\left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}\right] \tag{49}$$

将分母和分子写成多个条件概率相乘的形式：

$$= \mathbb{E}_q\left[\log \frac{\prod_{t=1}^T q(x_t|x_{t-1})}{p_\theta(x_T)\prod_{t=1}^T p_\theta(x_{t-1}|x_t)}\right] \tag{50}$$

$$= \mathbb{E}_q\left[-\log p_\theta(x_T) + \sum_{t=1}^T \log \frac{q(x_t|x_{t-1})}{p_\theta(x_{t-1}|x_t)}\right] \tag{51}$$

将 t=1 这一项单独拿出来，可以得到：

$$= \mathbb{E}_q\left[-\log p_\theta(x_T) + \sum_{t=2}^T \log \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{p_\theta(x_{t-1}|x_t)} + \log \frac{q(x_1|x_0)}{p_\theta(x_1)}\right] \tag{52}$$

$$\mathbb{E}_q\left[-\log p_\theta(x_T) + \sum_{t=2}^T \log\left(\frac{q(x_{t-1}\mid x_t, x_0)}{p_\theta(x_{t-1}\mid x_t)}\cdot\frac{q(x_t\mid x_0)}{q(x_{t-1}\mid x_0)}\right) + \log\left(\frac{q(x_1\mid x_0)}{p_\theta(x_0\mid x_1)}\right)\right] \tag{53}$$

$$= \mathbb{E}_q\left[-\log p_\theta(x_T) + \sum_{t=2}^T \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} + \sum_{t=2}^T \log \frac{q(x_t|x_0)}{q(x_{t-1}|x_0)} + \log \frac{q(x_1|x_0)}{p_\theta(x_1)}\right] \tag{54}$$

$$= \mathbb{E}_q\left[-\log p_\theta(x_T) + \sum_{t=2}^T \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} + \log \frac{q(x_T|x_0)}{q(x_1|x_0)} + \log \frac{q(x_1|x_0)}{p_\theta(x_1)}\right] \tag{55}$$

$$= \mathbb{E}_q \left[\log \frac{q(x_T|x_0)}{p_\theta(x_T)} + \sum_{t=2}^{T} \log \frac{q(x_{t-1}|x_t, x_0)}{p_\theta(x_{t-1}|x_t)} - \log p_\theta(x_0|x_1) \right] \tag{56}$$

$$= \mathbb{E}_q \left[\underbrace{D_{KL}(q(x_T \mid x_0) \parallel p_\theta(x_T))}_{L_T} + \sum_{t=2}^{T} \underbrace{D_{KL}(q(x_{t-1} \mid x_t, x_0) \parallel p_\theta(x_{t-1} \mid x_t))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0 \mid x_1)}_{L_0} \right] \tag{57}$$

从公式(57)中，我们可以看到 L_{VLB} 一共包含三个部分，第一个部分是不含参的，因为 $q(\cdot)$ 是不含参的，另外 $p_\theta(x_T)$ 是一个标准正态分布，也是不含参的。我们可以将 L_{t-1} 和 L_0 看成是一项。**（这一点很重要）论文中将 $p_\theta(x_{t-1} \mid x_t)$ 分布的方差设置为一个与 β 相关的常数，因此只训练的参数只存在于均值中。**

观察公式(57)的第二项，第一个分布 $q(x_{t-1} \mid x_t, x_0)$ 的分布是一个高斯分布，它的均值和方差我们在前面计算过，第二个分布 $p_\theta(x_{t-1} \mid x_t)$ 的分布我们在前面假设过，这也是一个高斯分布。我们参考公式(20)来计算两个高斯分布的 KL 散度：

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \left\| \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t) \right\|^2 \right] + C \tag{58}$$

所以，**我们的目标就是尽可能让 $\mu_\theta(\mathbf{x}_t, t)$ 和 $\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)$ 的距离足够小**:

$$L_t = \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{1}{2\|\Sigma_\theta(\mathbf{x}_t, t)\|_2^2} \|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, t)\|_2^2 \right] \tag{59}$$

$$= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{1}{2\|\Sigma_\theta\|_2^2} \left\| \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right) - \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) \right\|_2^2 \right] \tag{60}$$

$$= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\Sigma_\theta\|_2^2} \|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2 \right] \tag{61}$$

$$= \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\Sigma_\theta\|_2^2} \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\|_2^2 \right] \tag{62}$$

所以，公式(62)就是我们 loss 的最终形式（前面的系数可以忽略），它的意思是：**在我们训练的时候，我们有一个网络，它是以 x_o、α_t 和时刻 t 还有一个高斯噪声 ϵ，我们只需要让我们网络预测的噪声 ϵ_θ 和我们定义的高斯噪声 ϵ 越来越接近，就可以了**。

当我们训练好网络 ϵ_θ 后，我们就可以把它带入到公式(29)，然后利用重参数化技巧，得到 x_{t-1}、x_{t-2}，然后一直得到 x_0 了，也就是得到原图。

Algorithm 1 Training

> **repeat**
> $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
> $t \sim \text{Uniform}(\{1, \ldots, T\})$
> $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
> Take gradient descent step on
> $\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
> **until** converged

Algorithm 2 Sampling

> $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
> **for** $t = T, \ldots, 1$ **do**
> $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
> $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
> **end for**
> **return** $\mathbf{x}_0$

Figure 1: 相关伪代码

3 Question

3.1 为什么 DDPM 加噪声的幅度是不一样的呢？

DDPM 的前向过程是加噪的过程，将一张原图片不断的添加高斯噪声，直到最后变成一张纯噪声的图片。并且在添加噪声的过程中，噪声的强度是越来越大的，即 $\beta_1 < \beta_2 < ... < \beta_T$。那为什么幅度要越来越大呢？而不是幅度一致呢？

DDPM 的学习目标是学习一个去噪的过程，或者说是学习一个预测噪声的过程。我们在训练的时候，会初始化一个时间步 t，然后根据前向加噪的公式 $q(x_t \mid x_0)$ 来得到 x_t。如果说**加的噪声强度一致的话，那么模型在学习的时候，可能只能学习这一个强度的去噪过程，这可能导致模型在去噪任务上学得过于专一，缺乏对复杂情况的应对能力。**

3.2 在 DDPM 中，网络预测的是什么？

我们知道在 DDPM 的采样过程中，我们根据：

$$q(x_{t-1} \mid x_t, x_0) = q(x_t \mid x_{t-1}, x_0) \frac{q(x_{t-1} \mid x_0)}{q(x_t \mid x_0)} \tag{63}$$

又因为，DDPM 是服从马尔可夫假设的，因此 $q(x_t \mid x_{t-1}, x_0) = q(x_t \mid x_{t-1})$，然后根据之前小节的推导，我们可以得到：

$$p(x_{t-1} \mid x_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}\left(\mathbf{x}_t, \mathbf{x}_0\right), \tilde{\beta}_t \mathbf{I}\right) \tag{64}$$

在这个公式中，方差以及 x_t 都是已知的，只有 x_0 是未知的。因此，在这里有三种选择，第一种是直接预测 $\tilde{\boldsymbol{\mu}}\left(\mathbf{x}_t, \mathbf{x}_0\right)$，第二种是直接预测 x_0，第三种是预测噪声，然后根据噪声和 x_t 来反推得到 x_0，而在 DDPM 中，作者选择了第三种建模方式。