生成模型笔记

衔瑜*

2022 年 11 月 24 日

1 概率生成模型基本思想

我们可以将我们的训练集看作是对某个高维空间Ω中的一个随机向量X进行多次采样得到的结果，这个随机向量服从一个未知的概率分布P_r。那么我们可以通过X的样本集（即训练集）来学习获得一个参数化模型P_θ，此时学习获得的模型P_θ应当与P_r是近似的，之后通过学习到的模型P_θ再进行采样，则就能获得与训练集近似分布的生成结果，这便是生成网络的基本思想。

然而在实际操作中，由于待拟合的概率分布p_r往往非常复杂，神经网络通常很难直接拟合，且另一方面生成的结果很难量化评估其好坏，因此并不能简单地去构造一个生成网络来解决问题。

2 交叉熵

2.1 定义

对于分布为 $p(X)$ 的随机变量 X，我们用概率分布为 q 的编码对真实分布为 p 的信息 X 进行编码的长度我们就称为交叉熵，其表达式为

$$H(p,q)=E_{x\sim p(X)}[-\log q(X)]=\begin{cases}-\sum\limits_{x\in X}[p(x)\log q(x)] & ，\ x\sim p(X)\text{为离散概率分布}\\ -\int\limits_{x\in X}p(x)\log q(x)\,dx & ，\ x\sim p(X)\text{为连续概率分布}\end{cases}$$

在给定 p 的情况下，q 和 p 越接近，则交叉熵越小；反之则越大。

*Email: fish233yeah@163.com

3 KL散度

3.1 定义

KL散度（Kullback-Leibler Divergence）在概率模型中一般用于度量两个概率密度函数之间的“距离”，其定义为

$$KL[p(X)\|q(X)] = E_{x\sim p(X)}\left[\log\left(\frac{p(X)}{q(X)}\right)\right] = \begin{cases} \sum\limits_{x\in X}\left[p(x)\log\left(\frac{p(x)}{q(x)}\right)\right] & , \ x\sim p(X) \text{为离散概率分布} \\ \int\limits_{x\in X} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx & , \ x\sim p(X) \text{为连续概率分布} \end{cases}$$

可以证明KL散度必然大于等于0。注意到KL散度的定义中 $KL[p(X)\|q(X)]$ 关于 $p(X)$、 $q(X)$ 并不对称，即 $KL[p(X)\|q(X)] \neq KL[q(X)\|p(X)]$，因此KL散度不满足对称性，显然不是数学意义上的“度量”。

在优化问题中，常用 $p(X)$ 表示真实分布， $q(X)$ 表示一个用于拟合 $p(X)$ 的近似分布，在这种情形下，通常称 $KL[p(X)\|q(X)]$ 为前向KL散度（forward Kullback-Leibler Divergence），而称 $KL[q(X)\|p(X)]$ 为反向KL散度（reverse Kullback-Leibler Divergence）。

3.2 两类KL散度拟合效果定性分析

3.2.1 极小化前向KL散度代价下的拟合行为特性：寻找均值

前向KL的计算式中， $p(X)$ 和 $q(X)$ 在每个样本点 x 上的差异程度被 $p(X)$ 加权平均，我们基于此对前向KL的特性进行分析。

考虑随机变量 X 的子集 $X_0 = \{x|p(x) = 0\}$，由于 $p(X)$ 是前向KL公式中的权重系数，因此 X_0 中的元素实际上对前向KL的值没有任何影响。换言之，对任意 $x \in X_0$，无论 $p(X_0)$ 与 $q(X_0)$ 相差多大都对前向KL的计算结果毫无影响，因此前向KL值不受 $q(X)$ 在子集 $X_0 = \{x|p(x) = 0\}$ 上取值的影响。在极小化前向KL散度的过程中，每当 $p(X) = 0$ 时， $q(X)$ 就会被无视。从连续性角度推理，最小化前向KL散度倾向于忽视“ $q(X)$ 在满足 $p(X)$ 近似为0的随机变量取值集合上的拟合精度”，而去更努力的实现“ $q(X)$ 在满足 $p(X) > 0$ 的随机变量取值集合上的拟合精度”。前向KL散度的这种特性一般也被称为zero avoiding，原因是它倾向于避免在任何 $p(X) > 0$ 的位置 x 使得 $q(X) = 0$。最小化前向KL散度代价等价于对 $q(X)$ 中的参数 θ 的极大似然估计。图1左展示了使用前向KL散度代价拟合一个多峰（实际上是双峰）分布的效果示意图

3.2.2 极小化反向KL散度代价下的拟合行为特性：搜寻模态

在反向KL中，差异加权求和时的权重系数是 $q(X)$。此时， $p(X)$ 在子集 $X_0 = \{x|q(x) = 0\}$ 的取值不影响反向KL值的计算，而当 $q(X) > 0$ 时， $q(X)$ 与 $p(X)$ 的差异需要尽可能小以使得反向KL值尽可能小。因此，在以反向KL散度为代价的优化过程中， $q(X)$ 会更加趋向于要么取值接近0，要么去尽可能贴合 $p(X)$，表现即为 $q(X)$ 在拟合 $p(X)$ 时尽可能保持单一模态。

图1右展示了使用反向KL散度代价拟合一个多峰（实际上是双峰）分布的效果示意图

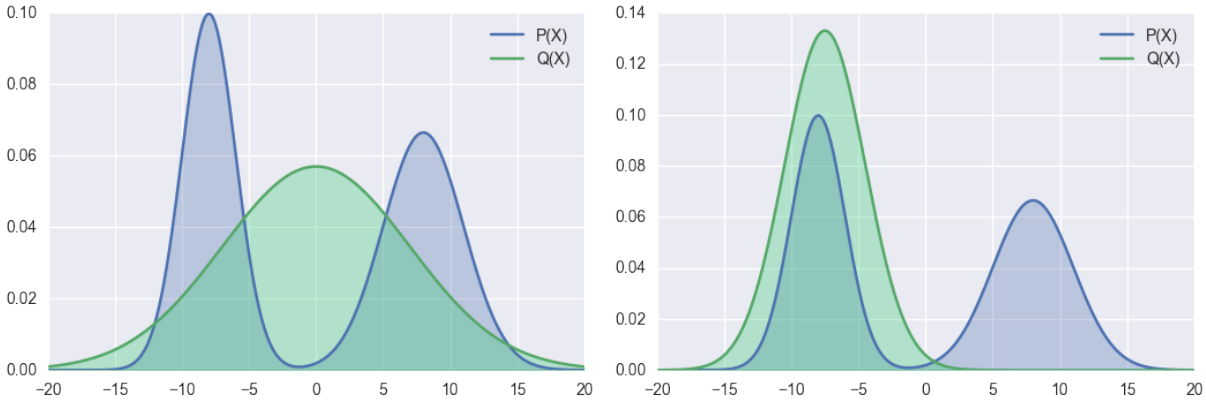

图 1: 左：前向KL代价拟合；右：反向KL代价拟合

4 VAE (Variational Autoencoders)

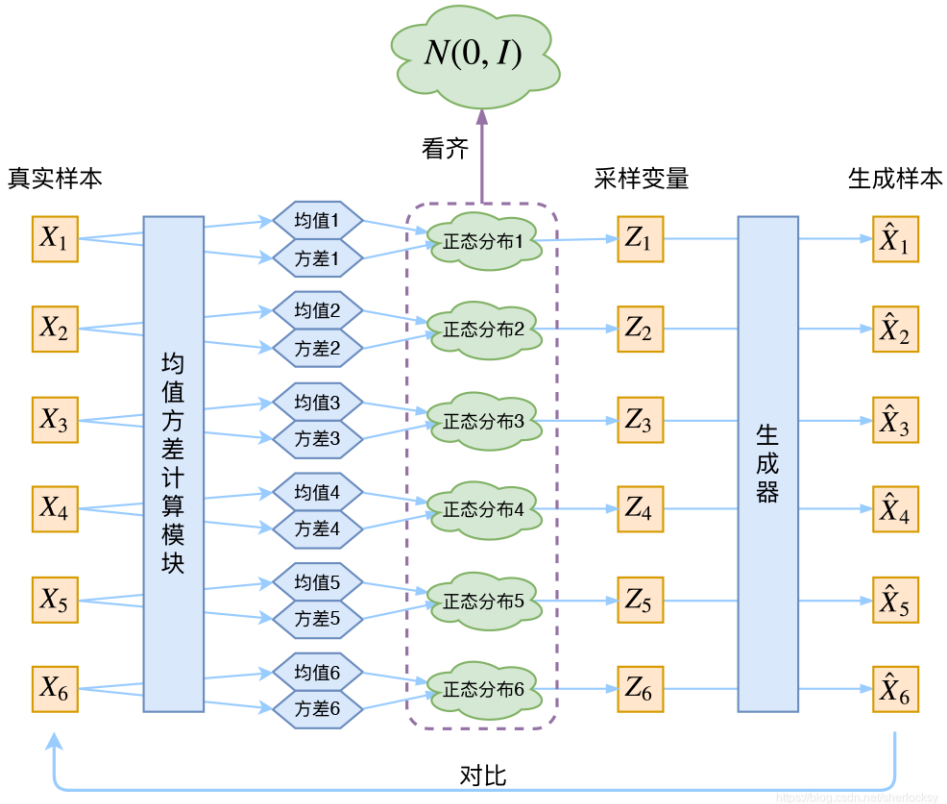

图 2: VAE网络结构

VAE整体可以看作encoder-decoder架构。对于一个输入的向量 x，它通常包含了许多冗余信息，我们要还原生成这个向量 x 实际上更需要的是描述其本质信息的隐变量 z，因此使用了encoder-decoder架构来设计网络，用encoder来进行 $x \to z$ 的编码过程，而decoder则用于 $z \to x$ 的生成过程。

首先考虑网络的decoder，即生成部分。如图3所示，由于目标概率密度函数 $p_r(X)$ 通常

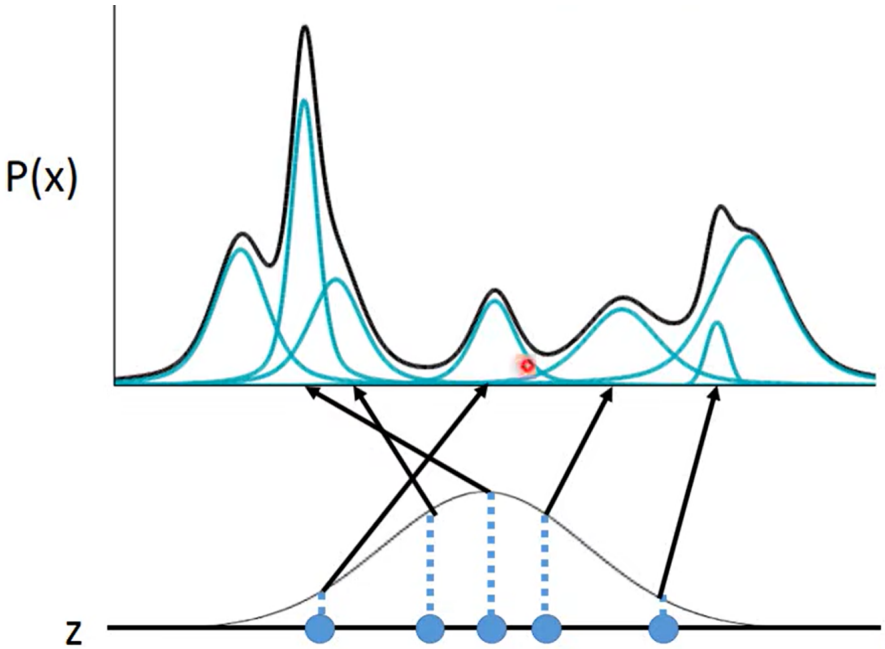

图 3: Z各点生成高斯分布叠加拟合$P(X)$

并不是一个上凸函数，它通常会具有多个极值，所以我们可以考虑使用多个高斯分布相叠加去拟合概率密度函数 $p_r(X)$，这个思想就是高斯混合模型的思想。我们将隐变量 Z 看作一个随机变量，我们对隐变量 Z 的每个可能的取值，都将其放入一个映射来得到一个均值 $\mu(z)$ 和一个标准差 $\sigma(z)$，此时我们便可以从这个隐变量的每个取值分别获得一个高斯分布 $N\left(\mu(z), \sigma^2(z)\right)$。因此，如果我们假设 Z 服从某个概率分布，那么对于概率密度函数 $p(X)$，我们有

$$p(x) = \int_z p(z)\, p(x|z)\, dz$$

其中 $p(X|Z)$ 服从高斯分布 $X|Z \sim N\left(\mu(Z), \sigma^2(Z)\right)$，这个 $p(X|Z)$ 便是 $z \to x$ 的结构，即VAE中的decoder结构。如果我们将 Z 的概率分布确定下来为某个已知分布，那么概率密度函数 $p(X)$ 就仅与映射 μ 和 σ 相关了。在VAE中我们假设 Z 的概率分布为 $Z \sim N(0, 1)$。

现在再来尝试推导出网络的encoder结构。由于我们的目标是在训练集上最大化 $\sum_x \log p(x)$，

我们假设 $Z|X$ 上有一个任意的概率密度函数 $q(Z|X)$，对 $\log p(X)$ 做变换我们有：

$$\begin{aligned}
\log p(x) &= \log p(x) \cdot \int_z q(z|x)\,dz \\
&= \int_z q(z|x) \log p(x)\,dz \\
&= \int_z q(z|x) \log\left(\frac{p(z,x)}{p(z|x)}\right) dz \\
&= \int_z q(z|x) \log\left(\frac{p(z,x)}{q(z|x)} \cdot \frac{q(z|x)}{p(z|x)}\right) dz \\
&= \int_z q(z|x) \log\left(\frac{p(z,x)}{q(z|x)}\right) dz + \int_z q(z|x) \log\left(\frac{q(z|x)}{p(z|x)}\right) dz \\
&= \int_z q(z|x) \log\left(\frac{p(x|z)\,p(z)}{q(z|x)}\right) dz + KL\left[q(z|x) \| p(z|x)\right]
\end{aligned}$$

我们记 $L_b = \int_z q(z|x) \log\left(\frac{p(x|z)p(z)}{q(z|x)}\right) dz$，则我们便得到了如下式子

$$\log p(x) = L_b + KL\left[q(z|x) \| p(z|x)\right] \geq L_b$$

此时很直观地，我们似乎只要调整 $p(x|z)$ 和 $q(z|x)$ 来将 L_b 最大化即可将 $\log p(x)$ 最大化了，但由于 $p(z|x)$ 与 $p(x|z)$ 相关联，在调整 $p(x|z)$ 的时候可能会导致 $KL\left[q(z|x) \| p(z|x)\right]$ 下降从而导致 $\log p(x)$ 反而下降，而由于 $q(z|x)$ 为任意概率密度函数，它变化时并不会导致 p 变化，因此我们考虑先把 $p(x|z)$ 固定（此时 $\log p(x) = \log\left(\int_z p(z)\,p(x|z)\,dz\right)$ 也会固定），仅调整 $q(z|x)$。显然，当 $q(z|x)$ 逼近 $p(z|x)$ 时， $KL\left[q(z|x) \| p(z|x)\right]$ 会趋近于0，此时 $\log p(x)$ 会趋近于 L_b，因此，在将 $q(z|x)$ 逼近 $p(z|x)$ 的情况下再通过调整 $p(z|x)$ 来增大 L_b 时， $\log p(x)$ 也会随之增大。而对于 L_b，我们有

$$\begin{aligned}
L_b &= \int_z q(z|x) \log\left(\frac{p(x|z)\,p(z)}{q(z|x)}\right) dz \\
&= \int_z q(z|x) \log p(x|z)\,dz + \int_z q(z|x) \log\left(\frac{p(z)}{q(z|x)}\right) dz \\
&= E_{Z\sim q(z|x)}\left[\log p(x|z)\right] - KL\left[q(z|x) \| p(z)\right]
\end{aligned}$$

此时我们便得到了我们最终所需的表达式：

$$\log p(x) = E_{Z\sim q(z|x)}\left[\log p(x|z)\right] - KL\left[q(z|x) \| p(z)\right] + KL\left[q(z|x) \| p(z|x)\right]$$

我们来逐个解析表达式中的三项内容：

1. $KL\left[q(z|x) \| p(z|x)\right]$
   在上文我们已经讲述过，我们要最大化 $\log p(x)$ 就会希望极小化 $KL\left[q(z|x) \| p(z|x)\right]$，这便意味着我们希望使用 $q(z|x)$ 去逼近 $p(z|x)$，即使用 $q(z|x)$ 来作为网络的encoder部分

2. $KL\left[q(z|x) \| p(z)\right]$
   要最大化 $\log p(x)$，则 $KL\left[q(z|x) \| p(z)\right]$ 就需要尽可能的小，而这项KL散度的直观含义

也很显然，即我们希望 $x \to z$ 的过程产生的 Z 的分布 $q(z|x)$ 尽可能与我们假设的 Z 的分布 $p(z)$ 近似，因此encoder部分还需要满足其概率分布为预先假设的 Z 的分布

3. $E_{Z\sim q(z|x)}[\log p(x|z)]$
   最大化 $\log p(x)$ 显然就会需要最大化期望 $E_{Z\sim q(z|x)}[\log p(x|z)]$，而最大化这项期望的含义也很直观：我们希望在给定encoder输出 $q(z|x)$ 的情况下，decoder的输出 $p(x|z)$ 的均值尽可能大。

现在对照网络结构图4，显然 $q(z|x)$ 对应的即为网络encoder部分，它需要去近似 Z 的分布，而根据我们对 $p(z)$ 的假设 $q(z|x)$ 将会趋近于一个标准正态分布，此时对 $q(z|x)$ 进行采样便可以得到 z，并传入decoder部分。网络的decoder部分则为最开始所说的 $p(x|z)$，即 $X|Z \sim N\left(\mu(Z), \sigma^2(Z)\right)$，用于使用高斯分布的输出去近似数据分布。

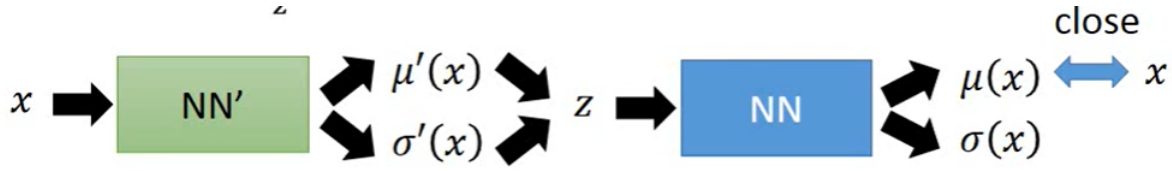

图 4: VAE算法结构

5 DDPM (Denoising Diffusion Probabilistic Models)

概率扩散模型的原理与VAE在某些地方有些相似。在DDPM中，我们将通过多次高斯扩散过程 $q(x_t|x_{t-1})$ 逐渐将源输入 X_0 转变为随机噪声 X_T，再从随机噪声 X_T 学习获得逆扩散过程 $p_\theta(x_{t-1}|x_t)$ 还原源输入 X_0。通过对扩散过程进行学习得到一系列马尔可夫逆扩散过程 $p_\theta(x_{t-1}|x_t)$ 后，我们便获得了从随机噪声到目标数据的条件分布 $\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$，如图5所示

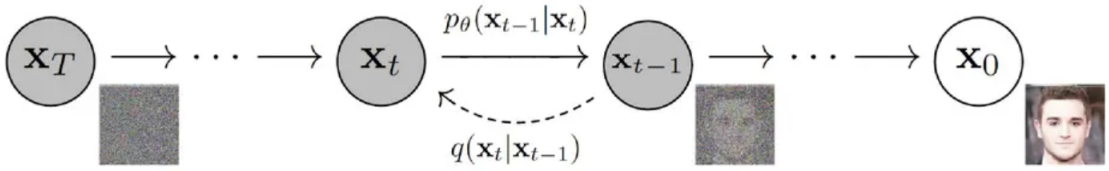

图 5: DDPM算法结构

我们先来讨论扩散过程。扩散过程实质上就是多次迭代向数据分布 $X_0 \sim q(x)$ 中添加高斯噪声的一个马尔可夫链过程，这个过程是固定的，且服从分布

$$q(X_t|X_{t-1}) = N\left(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I\right) \qquad q(X_{1:T}|X_0) = \prod_{t=1}^{T} q(X_t|X_{t-1})$$

其中 $\{\beta_t \in (0,1)\}_{t=1}^{T}$， I 为单位矩阵。由于这个过程是确定的，因此任意时刻的分布 $q(X_t)$ 完全可以由 X_0 及 $\beta_1, \beta_2, \cdots, \beta_t$ 确定。对 $X_t|X_{t-1}$，显然我们有

$$\frac{X_t|X_{t-1} - \sqrt{1-\beta_t}X_{t-1}}{\sqrt{\beta_t I}} \sim N(0, I)$$

设一系列 $\{Z \sim N(0,I)\}$ 以及设 $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$，此时我们从上式得到

$$X_t|X_{t-1} = \sqrt{\alpha_t}X_{t-1} + \sqrt{1-\alpha_t}Z_{t-1}$$

将上式中的 X_{t-1} 替换为条件分布 $X_{t-1}|X_{t-2}$，则

$$\begin{aligned}
X_t|X_{t-1}|X_{t-2} &= \sqrt{\alpha_t}\left(X_{t-1}|X_{t-2}\right) + \sqrt{1-\alpha_t}Z_{t-1} \\
X_t|X_{t-1}, X_{t-2} &= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}X_{t-2} + \sqrt{1-\alpha_{t-1}}Z_{t-2}\right) + \sqrt{1-\alpha_t}Z_{t-1} \\
X_t|X_{t-1}, X_{t-2} &= \sqrt{\alpha_t\alpha_{t-1}}X_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}Z_{t-2} + \sqrt{1-\alpha_t}Z_{t-1} \\
X_t|X_{t-1}, X_{t-2} &= \sqrt{\alpha_t\alpha_{t-1}}X_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}Z_{t-1,t-2}
\end{aligned}$$

以此类推，最终我们可以得到

$$X_t|X_{t-1}, \cdots, X_0 = \sqrt{\bar{\alpha}_t}X_0 + \sqrt{1-\bar{\alpha}_t}Z_{t-1,t-2,\cdots,0}$$

即我们有

$$q\left(X_t|X_{0:(t-1)}\right) = N\left(X_t;\ \sqrt{\bar{\alpha}_t}X_0, (1-\bar{\alpha}_t)I\right)$$

由于 $N\left(X_t;\ \sqrt{\bar{\alpha}_t}X_0, (1-\bar{\alpha}_t)I\right)$ 仅与 X_0 有关，因此我们有

$$q\left(X_t|X_0\right) = N\left(X_t;\ \sqrt{\bar{\alpha}_t}X_0, (1-\bar{\alpha}_t)I\right)$$

且有

$$q\left(X_t|X_{t-1}\right) = q\left(X_t|X_{0:(t-1)}\right) = q\left(X_t|X_0\right)$$

显然，当 $\bar{\alpha}_t \to 0$ 时， $q(X_t|X_0)$ 趋近于服从标准正态分布。在原文中，作者将 $\{\beta_t\}_{t=1}^{T}$ 设置为了线性增加的序列，即 $\beta_1 < \beta_2 < \cdots < \beta_T$。

接下来我们讨论逆扩散过程。第一步我们先推导一个我们后面会需要用到的后验条件分

布：

$$
\begin{aligned}
q\left(X_{t-1}|X_t,X_0\right) &= \frac{q\left(X_t,X_{t-1},X_0\right)}{q\left(X_t,X_0\right)} \\
&= \frac{q\left(X_t|X_{t-1},X_0\right)q\left(X_{t-1},X_0\right)}{q\left(X_t,X_0\right)} \\
&= q\left(X_t|X_{t-1}\right)\frac{q\left(X_{t-1}|X_0\right)}{q\left(X_t|X_0\right)} \\
&= \frac{1}{\sqrt{2\pi}\cdot\frac{\sigma_{q(X_t|X_{t-1})}\sigma_{q(X_{t-1}|X_0)}}{\sigma_{q(X_t|X_0)}}}\cdot e^{-\frac{1}{2}\left[\frac{\left(X_t-\sqrt{\alpha_t}X_{t-1}\right)^2}{\beta_t}+\frac{\left(X_{t-1}-\sqrt{\bar{\alpha}_{t-1}}X_0\right)^2}{1-\bar{\alpha}_{t-1}}-\frac{\left(X_t-\sqrt{\bar{\alpha}_t}X_0\right)^2}{1-\bar{\alpha}_t}\right]} \\
&= \frac{1}{\sqrt{2\pi}\cdot\frac{\sigma_{q(X_t|X_{t-1})}\sigma_{q(X_{t-1}|X_0)}}{\sigma_{q(X_t|X_0)}}}\cdot e^{-\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(X_t,X_0)\right]} \\
&= \frac{1}{\sqrt{2\pi}\cdot\frac{\sigma_{q(X_t|X_{t-1})}\sigma_{q(X_{t-1}|X_0)}}{\sigma_{q(X_t|X_0)}}}\cdot e^{-\frac{1}{2}\left[\frac{\left(X_{t-1}-\frac{\frac{\sqrt{\alpha_t}}{\beta_t}X_t+\frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}}X_0}{\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}}\right)^2}{\frac{1}{\frac{\alpha_t}{\beta_t}+\frac{1}{1-\bar{\alpha}_{t-1}}}}\right]+C'(X_t,X_0)} \\
&= \frac{1}{\sqrt{2\pi}\cdot\frac{\sigma_{q(X_t|X_{t-1})}\sigma_{q(X_{t-1}|X_0)}}{\sigma_{q(X_t|X_0)}}}e^{C'(X_t,X_0)}\cdot e^{-\frac{1}{2}\left[\frac{\left(X_{t-1}-\left(\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\right)\right)^2}{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}\right]} \\
&= \frac{1}{\sqrt{2\pi}\cdot\sigma'}\cdot e^{-\frac{1}{2}\left[\frac{\left(X_{t-1}-\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}X_t+\frac{\sqrt{\bar{\alpha}_t-1}\cdot\beta_t}{1-\bar{\alpha}_t}X_0\right)\right)^2}{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}\right]}
\end{aligned}
$$

其中， $C\left(X_t,X_0\right)$ 和 $C'\left(X_t,X_0\right)$ 均为某个仅与 X_0 和 X_t 有关的函数, $\sigma'=\frac{\sigma_{q(X_t|X_{t-1})}\cdot\sigma_{q(X_{t-1}|X_0)}}{\sigma_{q(X_t|X_0)}\cdot e^{C'(X_t,X_0)}}$。显然，对均值为 $\tilde{\mu}_t\left(X_t,X_0\right)=\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}X_t+\frac{\sqrt{\bar{\alpha}_t-1}\cdot\beta_t}{1-\bar{\alpha}_t}X_0$，方差为 $\tilde{\beta}_tI=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\cdot I$ 的正态分布 $N\left(X_{t-1};\tilde{\mu}_t\left(X_t,X_0\right),\tilde{\beta}_tI\right)$，我们有

$$
\frac{q\left(X_{t-1}|X_t,X_0\right)}{f_{N\left(\tilde{\mu}_t(X_t,X_0),\tilde{\beta}_tI\right)}\left(X_{t-1}\right)}=\frac{\sqrt{\tilde{\beta}_t}}{\sigma'}
$$

而由于

$$
\int q\left(X_{t-1}|X_t,X_0\right)dx=1=\int f_{N\left(\tilde{\mu}_t(X_t,X_0),\tilde{\beta}_tI\right)}\left(X_{t-1}\right)dx
$$

因此显然有 $\sigma'=\sqrt{\tilde{\beta}_t}$，即后验分布 $q\left(X_{t-1}|X_t,X_0\right)$ 服从正态分布 $N\left(X_{t-1};\tilde{\mu}_t\left(X_t,X_0\right),\tilde{\beta}_tI\right)$。

现在我们假设逆扩散过程也是一个马尔可夫链过程，且有

$$
p_\theta\left(X_{0:(T-1)}|X_T\right)=\prod_{t=1}^{T}p_\theta\left(X_{t-1}|X_t\right)
$$

那么对于其对数似然函数，我们有

$$
\begin{aligned}
-\log p_\theta(X_0) &\leq -\log p_\theta(X_0) + KL\left[q(X_{1:T}|X_0)\|p_\theta(X_{1:T}|X_0)\right] \\
&= -\log p_\theta(X_0) + E_{x_{1:T}\sim q(X_{1:T}|X_0)}\left[\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{1:T}|X_0)}\right)\right] \\
&= -\log p_\theta(X_0) + E_{x_{1:T}\sim q(X_{1:T}|X_0)}\left[\log\left(\frac{q(X_{1:T}|X_0)}{\frac{p_\theta(X_{0:T})}{p_\theta(X_0)}}\right)\right] \\
&= -\log p_\theta(X_0) + E_{x_{1:T}\sim q(X_{1:T}|X_0)}\left[\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{0:T})}\right)\right] + E_{x_{1:T}\sim q(X_{1:T}|X_0)}\left[\log\left(p_\theta(X_0)\right)\right] \\
&= E_{x_{1:T}\sim q(X_{1:T}|X_0)}\left[\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{0:T})}\right)\right]
\end{aligned}
$$

将对数似然转换为交叉熵，我们有

$$
\begin{aligned}
E_{x\sim q(X_0)}\left[-\log p_\theta(X_0)\right] &\leq E_{x\sim q(X_0)}\left\{E_{x_{1:T}\sim q(X_{1:T}|X_0)}\left[\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{0:T})}\right)\right]\right\} \\
&= E_{x\sim q(X_0)}\left\{\int q(X_{1:T}|X_0)\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{0:T})}\right)dX_{1:T}\right\} \\
&= \int q(X_0)\int q(X_{1:T}|X_0)\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{0:T})}\right)dX_{1:T}dX_0 \\
&= \int q(X_{0:T})\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{0:T})}\right)dX_{0:T} \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{0:T})}\right)\right]
\end{aligned}
$$

设 $L_{VLB} = E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q(X_{1:T}|X_0)}{p_\theta(X_{0:T})}\right)\right]$，那么要最小化交叉熵 $E_{x\sim q(X_0)}\left[-\log p_\theta(X_0)\right]$ 可以考虑直接

最小化 L_{VLB}。继续推导 L_{VLB}，我们有

$$
\begin{aligned}
L_{VLB} &= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_{1:T}|X_0\right)}{p_\theta\left(X_{0:T}\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{\prod_{t=1}^{T} q\left(X_t|X_{t-1}\right)}{p_\theta\left(X_T\right)\prod_{t=1}^{T} p_\theta\left(X_{t-1}|X_t\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[-\log p_\theta\left(X_T\right)+\sum_{t=1}^{T}\log\left(\frac{q\left(X_t|X_{t-1}\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[-\log p_\theta\left(X_T\right)+\log\left(\frac{q\left(X_1|X_0\right)}{p_\theta\left(X_0|X_1\right)}\right)+\sum_{t=2}^{T}\log\left(\frac{q\left(X_t|X_{t-1},X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[-\log p_\theta\left(X_T\right)+\log\left(\frac{q\left(X_1|X_0\right)}{p_\theta\left(X_0|X_1\right)}\right)+\sum_{t=2}^{T}\log\left(\frac{q\left(X_t,X_{t-1},X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)\cdot q\left(X_{t-1},X_0\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[-\log p_\theta\left(X_T\right)+\log\left(\frac{q\left(X_1|X_0\right)}{p_\theta\left(X_0|X_1\right)}\right)+\sum_{t=2}^{T}\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)\cdot q\left(X_t,X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)\cdot q\left(X_{t-1},X_0\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[-\log p_\theta\left(X_T\right)+\log\left(\frac{q\left(X_1|X_0\right)}{p_\theta\left(X_0|X_1\right)}\right)+\sum_{t=2}^{T}\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)}\cdot\frac{q\left(X_t|X_0\right)}{q\left(X_{t-1}|X_0\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[-\log p_\theta\left(X_T\right)+\log\left(\frac{q\left(X_1|X_0\right)}{p_\theta\left(X_0|X_1\right)}\right)+\sum_{t=2}^{T}\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)+\sum_{t=2}^{T}\log\left(\frac{q\left(X_t|X_0\right)}{q\left(X_{t-1}|X_0\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[-\log p_\theta\left(X_T\right)+\sum_{t=2}^{T}\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)+\log\left(\frac{q\left(X_T|X_0\right)}{p_\theta\left(X_0|X_1\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_T|X_0\right)}{p_\theta\left(X_T\right)}\right)+\sum_{t=2}^{T}\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)-\log p_\theta\left(X_0|X_1\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_T|X_0\right)}{p_\theta\left(X_T\right)}\right)\right]+\sum_{t=2}^{T}E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)\right]-\log p_\theta\left(X_0|X_1\right)
\end{aligned}
$$

最后一步的 $E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q(X_T|X_0)}{p_\theta(X_T)}\right)\right]$ 以及 $T-1$ 项 $E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q(X_{t-1}|X_t,X_0)}{p_\theta(X_{t-1}|X_t)}\right)\right]$ 它们并不是KL散度，因此在这里并没有写成KL散度表示形式。记

$$
\begin{aligned}
L_T &= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_T|X_0\right)}{p_\theta\left(X_T\right)}\right)\right] \\
L_{t-1} &= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)\right] \\
L_0 &= -\log p_\theta\left(X_0|X_1\right)
\end{aligned}
$$

由于最后一步的 $E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q(X_T|X_0)}{p_\theta(X_T)}\right)\right]$ 以及 $T-1$ 项 $E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q(X_{t-1}|X_t,X_0)}{p_\theta(X_{t-1}|X_t)}\right)\right]$ 并不是KL散度，因此我们并不能确定其优化下界是否存在，以及存在的话下界为多少，因此在讨论损失函数优化目标之前我们还需要先证明一下各个 $L_i, i=0,1,\cdots,T$ 的值的取值范围。

1. L_T

$$\begin{aligned}
L_T &= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_T|X_0\right)}{p_\theta\left(X_T\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_T|X_{0:(T-1)}\right)}{p_\theta\left(X_T\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_{0:T}\right)}{p_\theta\left(X_T\right)}\right)+\log\left(\frac{1}{q\left(X_{0:(T-1)}\right)}\right)\right] \\
&= KL\left[q\left(X_{0:T}\right)\|p_\theta\left(X_T\right)\right]+E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{1}{q\left(X_{0:(T-1)}\right)}\right)\right]
\end{aligned}$$

由于 $KL\left[q\left(X_{0:T}\right)\|p_\theta\left(X_T\right)\right]$ 和 $E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{1}{q\left(X_{0:(T-1)}\right)}\right)\right]$ 均大于等于0，因此 L_T 大于等于0；又由于当 $q\left(X_T|X_0\right)=p_\theta\left(X_T\right)$ 时 L_T 等于0，因此 L_T 的优化下界存在且为0。

2. L_{t-1}

$$\begin{aligned}
L_{t-1} &= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)\right] \\
&= E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_{0:T}\right)}{p_\theta\left(X_{t-1}|X_t\right)}\right)+\log\left(\frac{q\left(X_{t-1}|X_t,X_0\right)}{q\left(X_{0:T}\right)}\right)\right] \\
&= KL\left[q\left(X_{0:T}\right)\|p_\theta\left(X_{t-1}|X_t\right)\right]+E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_t|X_{t-1}\right)q\left(X_{t-1}|X_0\right)}{q\left(X_{0:T}\right)q\left(X_t|X_0\right)}\right)\right] \\
&= KL\left[q\left(X_{0:T}\right)\|p_\theta\left(X_{t-1}|X_t\right)\right]+E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_t|X_{0:(t-1)}\right)q\left(X_{t-1}|X_{0:(t-2)}\right)}{q\left(X_{0:T}\right)q\left(X_t|X_{0:(t-1)}\right)}\right)\right] \\
&= KL\left[q\left(X_{0:T}\right)\|p_\theta\left(X_{t-1}|X_t\right)\right]+E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{q\left(X_{0:(t-1)}\right)}{q\left(X_{0:T}\right)q\left(X_{0:(t-2)}\right)}\right)\right] \\
&= KL\left[q\left(X_{0:T}\right)\|p_\theta\left(X_{t-1}|X_t\right)\right]+E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{1}{q\left(X_{t:T}|X_{0:(t-1)}\right)q\left(X_{0:(t-2)}\right)}\right)\right]
\end{aligned}$$

由于 $KL\left[q\left(X_{0:T}\right)\|p_\theta\left(X_{t-1}|X_t\right)\right]$ 和 $E_{x_{0:T}\sim q(X_{0:T})}\left[\log\left(\frac{1}{q\left(X_{t:T}|X_{0:(t-1)}\right)q\left(X_{0:(t-2)}\right)}\right)\right]$ 均大于等于0，因此 L_{t-1} 大于等于0；又由于当 $q\left(X_{t-1}|X_t,X_0\right)=p_\theta\left(X_{t-1}|X_t\right)$ 时 L_{t-1} 等于0，因此 L_{t-1} 的优化下界存在且为0。

3. L_0

显然 L_0 的优化下界存在且为0。

因此，根据上述推导过程，要最小化 L_{VLB} 我们就需要最小化各个 $L_i, i=0,1,\cdots,T$，且希望模型逆扩散过程的每一步都去拟合对应的扩散过程的反向概率分布，通过对扩散过程的学习来得到逆扩散的分布，即让 $p_\theta\left(X_T\right)$ 学习 $q\left(X_T|X_0\right)$，让 $p_\theta\left(X_{t-1}|X_t\right)$ 学习 $q\left(X_{t-1}|X_t,X_0\right)$，最后让 $p_\theta\left(X_0|X_1\right)$ 逼近1。而又由于我们证明了 $q\left(X_T|X_0\right)$ 近似服从标准正态分布，且每一步 $q\left(X_{t-1}|X_t,X_0\right)$ 均服从正态分布，因此逆扩散过程的每一步也服从正态分布，因此我们可以假设

$$p_\theta\left(X_T\right)=N\left(X_T;0,1\right)\qquad p_\theta\left(X_{t-1}|X_t\right)=N\left(X_{t-1};\mu_\theta\left(X_t,t\right),\Sigma_\theta\left(X_t,t\right)\right)$$

来对生成过程进行拟合。至此，网络结构设计和损失函数设计均推导完毕。

参考文献

[1] 邱锡鹏. 神经网络与深度学习. 机械工业出版社. https://nndl.github.io/. 2020

[2] 工具人66号. 进阶详解KL散度. 知乎. 2022

[3] Hung-yi Lee. ML Lecture 18: Unsupervised Learning - Deep Generative Model (Part II). Youtube. 2016

[4] deep_thoughts. Probabilistic Diffusion Model概率扩散模型理论与完整PyTorch代码详细解读. bilibili. 2022