父条目： Understanding Diffusion Models: A Unified Perspective

笔记

生成模型

在生成模型中，往往寻求低维的表示，而不是高维的；因为除非有先验，否则效果很差；而且低维可以看作一种压缩形式，表述语义信息；

前置知识

把隐变量和data视为一个联合分布p(x, z)，有两种方式可以把对于所有x的似然p(x)最大化：

1.可以显式地把z边缘化：

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}$$

2.也可以用概率的链式法则

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}$$

显然地，要么积分，要么得有个真实值p(z|x)，所以直接计算p(x)很困难；但可以利用这两个方程推出一个叫证据下界ELBO的东西，在最好情况下，这两者等价；

ELBO

Evidence Lower Bound，定义如下：

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right]$$

显然，可以写出证据p(x)与ELBO的关系

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right]$$

$q_\varphi(z|x)$是一个具有参数φ的近似的变分分布，直观上是一个估计给定参数x的真实分布，要尽可能近似真实的后验p(z|x)；

从等式1可以推导

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} && \text{(Apply Equation 1)} \\
&= \log \int \frac{p(\boldsymbol{x}, \boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} && \text{(Multiply by } 1 = \tfrac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\text{)} \\
&= \log \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(Definition of Expectation)} \\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(Apply Jensen's Inequality)}
\end{aligned}$$

但这不够直观，不妨再从等式2看一下：

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log p(\boldsymbol{x}) \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) dz && \text{(Multiply by } 1 = \textstyle\int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}\text{)} && (9) \\
&= \int q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})(\log p(\boldsymbol{x})) dz && \text{(Bring evidence into integral)} && (10) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x})\right] && \text{(Definition of Expectation)} && (11) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(Apply Equation 2)} && (12) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(Multiply by } 1 = \tfrac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\text{)} && (13) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(Split the Expectation)} && (14) \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] + D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x})) && \text{(Definition of KL Divergence)} && (15) \\
&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{(KL Divergence always} \geq 0\text{)} && (16)
\end{aligned}$$

这相当于把两者之间的具体差值求出来了，这个KL散度在等式1用琴森不等式时被放缩掉了；

为什么要优化ELBO？

1.KL项非负，所以ELBO的值不会超过证据；

2.在引入想要建模的隐变量z后，希望优化变分后验$q_\varphi(z|x)$，即最小化KL散度，来精确匹配真正的后验分布p(z|x)。但显然，没有ground truth p(z|x)，所以无法直接最小化KL散度。但式15里的似然都是关于φ的常数，不依赖于φ。ELBO+KL散度是定值，所以对ELBO的最大化就是对KL散度的最小化，所以优化ELBO的过程就是让近似后验接近真实后验的过程；

Variational Autoencoders

即VAE，直接最大化ELBO

变分方法

$$\begin{aligned}\mathbb{E}_{q_{\phi}(z|x)}\left[\log\frac{p(x,z)}{q_{\phi}(z|x)}\right] &= \mathbb{E}_{q_{\phi}(z|x)}\left[\log\frac{p_{\theta}(x|z)p(z)}{q_{\phi}(z|x)}\right] && \text{(Chain Rule of Probability)} \quad (17)\\ &= \mathbb{E}_{q_{\phi}(z|x)}\left[\log p_{\theta}(x|z)\right] + \mathbb{E}_{q_{\phi}(z|x)}\left[\log\frac{p(z)}{q_{\phi}(z|x)}\right] && \text{(Split the Expectation)} \quad (18)\\ &= \underbrace{\mathbb{E}_{q_{\phi}(z|x)}\left[\log p_{\theta}(x|z)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q_{\phi}(z|x)\parallel p(z))}_{\text{prior matching term}} && \text{(Definition of KL Divergence)} \quad (19)\end{aligned}$$

在这种情况下，学习了一个中间的bottleneck分布$q_\varphi(z|x)$，把输入转为可能隐变量的分布，类似编码器；还学习了一个确定性函数$p_\theta(x|z)$，把隐变量z转化为观测x，类似解码器；

显然可以直观解释：前者是reconstruction项，描述了变分分布与ground truth分布的相似性；后者是prior matching term，越小说明越相似；

这个方法需要联合优化两个参数φ和θ。通常如下初始化

$$\begin{aligned}q_{\phi}(z|x) &= \mathcal{N}(z;\mu_{\phi}(x),\sigma_{\phi}^{2}(x)\mathbf{I})\\ p(z) &= \mathcal{N}(z;\mathbf{0},\mathbf{I})\end{aligned}$$

然后计算重建项的蒙特卡洛近似和KL散度的解析解

$$\arg\max_{\phi,\theta}\mathbb{E}_{q_{\phi}(z|x)}\left[\log p_{\theta}(x|z)\right] - D_{\mathrm{KL}}(q_{\phi}(z|x)\parallel p(z)) \approx \arg\max_{\phi,\theta}\sum_{l=1}^{L}\log p_{\theta}(x|z^{(l)}) - D_{\mathrm{KL}}(q_{\phi}(z|x)\parallel p(z)) \quad (22)$$

这个$z^{(l)}$是从$q_\varphi(z|x)$中对每个样本x计算的结果抽样得到的；需要注意，每个$z^{(l)}$都是随机抽样的，不可微，但可以重参数化解决；

重参数化

重要技巧，参见另外的笔记

主要思路是把原先随机采样的z的"随机性"分散到了一个确定的分布中，如高斯分布，从而可以进行反向求导；

$$z = \mu_{\phi}(x) + \sigma_{\phi}(x)\odot\epsilon \quad \text{with } \epsilon\sim\mathcal{N}(\epsilon;\mathbf{0},\mathbf{I})$$

Hierarchical Variational Autoencoders

HVAE，相较于VAE有更多层次，隐变量本身也可以是由更高层次、更抽象的隐变量构成；

考虑一个特殊情况：Markovian HVAE，它的生成过程是一个马尔可夫链；

同样的，可以有：

$$p(x,z_{1:T}) = p(z_T)p_{\theta}(x|z_1)\prod_{t=2}^{T}p_{\theta}(z_{t-1}|z_t)$$

$$q_{\phi}(z_{1:T}|x) = q_{\phi}(z_1|x)\prod_{t=2}^{T}q_{\phi}(z_t|z_{t-1})$$

类似地，ELBO为

$$\begin{aligned}\log p(x) &= \log\int p(x,z_{1:T})dz_{1:T} && \text{(Apply Equation 1)}\\ &= \log\int\frac{p(x,z_{1:T})q_{\phi}(z_{1:T}|x)}{q_{\phi}(z_{1:T}|x)}dz_{1:T} && \text{(Multiply by } 1 = \tfrac{q_{\phi}(z_{1:T}|x)}{q_{\phi}(z_{1:T}|x)}\text{)}\\ &= \log\mathbb{E}_{q_{\phi}(z_{1:T}|x)}\left[\frac{p(x,z_{1:T})}{q_{\phi}(z_{1:T}|x)}\right] && \text{(Definition of Expectation)}\\ &\geq \mathbb{E}_{q_{\phi}(z_{1:T}|x)}\left[\log\frac{p(x,z_{1:T})}{q_{\phi}(z_{1:T}|x)}\right] && \text{(Apply Jensen's Inequality)}\end{aligned}$$

代入联合分布p和后验q

$$\mathbb{E}_{q_{\phi}(z_{1:T}|x)}\left[\log\frac{p(x,z_{1:T})}{q_{\phi}(z_{1:T}|x)}\right] = \mathbb{E}_{q_{\phi}(z_{1:T}|x)}\left[\log\frac{p(z_T)p_{\theta}(x|z_1)\prod_{t=2}^{T}p_{\theta}(z_{t-1}|z_t)}{q_{\phi}(z_1|x)\prod_{t=2}^{T}q_{\phi}(z_t|z_{t-1})}\right]$$

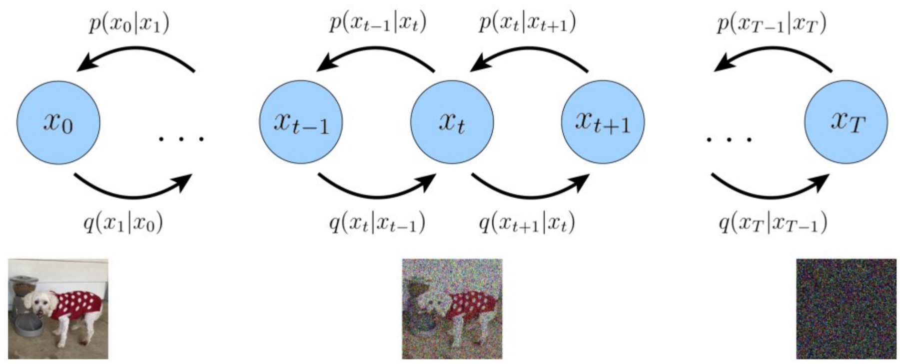
Variational Diffusion Models

VDM是MHVAE的一个特殊情况：

1.隐变量维度等于数据维度

2.隐编码器不在时刻t维度上学习，被预定义为一个线性的高斯模型，即一个以前一个时刻的输出为中心的分布

3.隐编码器的高斯参数随时间变化，使得隐编码在最终时刻T处的分布为标准高斯分布；



由第一个假设：

VDM可以被重写为

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

由第二个假设：

由于不同层之间的高斯编码器是预先设定好的线性高斯模型，所以设置均值$\mu_t(x_t)=\sqrt{\alpha_t}x_{t-1}$，方差$\Sigma_t(x_t)=(1-\alpha_t)I$，这样隐变量就能在编码过程中保持方差。

编码器的转换过程可以如下表述：

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

由第三个假设：

αt是固定的，且最终的隐变量p(xT)是一个标准高斯分布，所以可以把联合分布重写为：

$$p(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$$

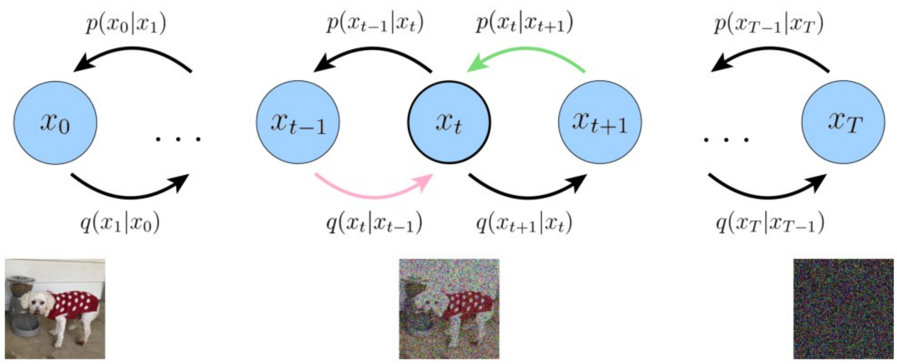
$$\text{where,}$$

$$p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$$

需要注意，此时$q(x_t|x_{t-1})$是不受参数φ的影响的，因为整个encoder都是用定义好的高斯分布来建模的。所以只需要学习$p_\theta(x_t|x_{t-1})$，这样就可以生成新的数据了

$$
\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} && (34)\\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} && (35)\\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] && (36)\\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] && (37)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (38)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (39)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \prod_{t=1}^{T-1} p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (40)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (41)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (42)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (43)\\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right] + \mathbb{E}_{q(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1})} \right] + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] && (44)\\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_{\theta}(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T)) \right]}_{\text{prior matching term}} && (45)\\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_{\theta}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})) \right]}_{\text{consistency term}}
\end{aligned}
$$

这样把证据分为了3个部分：reconstruction项，与VAE类似；prior matching项，无需训练；和consistency项，它希望加的噪声和去的噪声应该相匹配；



所有步骤都是在计算期望，因此可以用蒙特卡洛近似；

改进1

但是，对于第三项而言，每次会对两个随机变量$\{x_{t-1}, x_{t+1}\}$进行估计，偏差很大；

注意到显然$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0)$，用贝叶斯定理有

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

于是

$$
\begin{aligned}
\log p(\boldsymbol{x}) &\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] & (47)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] & (48)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] & (49)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\prod_{t=2}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)\prod_{t=2}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right] & (50)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right] & (51)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}\right] & (52)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\cancel{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}}{\cancel{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}}\right] & (53)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}} + \log \frac{\cancel{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \log \prod_{t=2}^{T} \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] & (54)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \sum_{t=2}^{T}\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] & (55)\\
&= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T}\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] & (56)\\
&= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \sum_{t=2}^{T}\mathbb{E}_{q(\boldsymbol{x}_t,\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] & (57)\\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}} & (58)
\end{aligned}
$$

这样，每次估计只有一个随机变量；

类似的，证据里面也有三项：reconstruction项，与VAE类似；prior matching项，无需训练，且在此假设下应为0；denoising matching项，它定义了一个期望的去噪过程p，作为真实的去噪过程$q(x_{t-1}|x_t, x_0)$的近似；

优化开销

显然，优化过程开销主要还是在求和项（第三项）上。在VDM中，可以用Gaussian transition假设来优化，首先有

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

此时，$q(x_t|x_{t-1}, x_0)=q(x_t|x_{t-1})$，为高斯分布；

接下来需要计算$q(x_t|x_0)$和$q(x_{t-1}|x_0)$，利用重参数化技巧，变量可以如下转化

$$
\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\mathbf{I})
$$

$$
\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\mathbf{I})
$$

于是，代入后有

$$
\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}^{*} & (61)\\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}^{*}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}^{*} & (62)\\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2}^{*} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}^{*} & (63)\\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\sqrt{\alpha_t-\alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\boldsymbol{\epsilon}_{t-2} & (64)\\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t-\alpha_t\alpha_{t-1}+1-\alpha_t}\boldsymbol{\epsilon}_{t-2} & (65)\\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} & (66)\\
&= \ldots & (67)\\
&= \sqrt{\prod_{i=1}^{t}\alpha_i}\boldsymbol{x}_0 + \sqrt{1-\prod_{i=1}^{t}\alpha_i}\boldsymbol{\epsilon}_0 & (68)\\
&= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 & (69)\\
&\sim \mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,(1-\bar{\alpha}_t)\mathbf{I}) & (70)
\end{aligned}
$$

也可以用另一种形式推导

$$
\begin{aligned}
q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} && (71)\\
&= \frac{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\alpha_t}\boldsymbol{x}_{t-1},(1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1};\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0,(1-\bar\alpha_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar\alpha_t}\boldsymbol{x}_0,(1-\bar\alpha_t)\mathbf{I})} && (72)\\
&\propto \exp\left\{-\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{2(1-\alpha_t)}+\frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar\alpha_{t-1})}-\frac{(\boldsymbol{x}_t-\sqrt{\bar\alpha_t}\boldsymbol{x}_0)^2}{2(1-\bar\alpha_t)}\right]\right\} && (73)\\
&= \exp\left\{-\frac{1}{2}\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{1-\alpha_t}+\frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0)^2}{1-\bar\alpha_{t-1}}-\frac{(\boldsymbol{x}_t-\sqrt{\bar\alpha_t}\boldsymbol{x}_0)^2}{1-\bar\alpha_t}\right]\right\} && (74)\\
&= \exp\left\{-\frac{1}{2}\left[\frac{(-2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}+\alpha_t\boldsymbol{x}_{t-1}^2)}{1-\alpha_t}+\frac{(\boldsymbol{x}_{t-1}^2-2\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0)}{1-\bar\alpha_{t-1}}+C(\boldsymbol{x}_t,\boldsymbol{x}_0)\right]\right\} && (75)\\
&\propto \exp\left\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}}{1-\alpha_t}+\frac{\alpha_t\boldsymbol{x}_{t-1}^2}{1-\alpha_t}+\frac{\boldsymbol{x}_{t-1}^2}{1-\bar\alpha_{t-1}}-\frac{2\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0}{1-\bar\alpha_{t-1}}\right]\right\} && (76)\\
&= \exp\left\{-\frac{1}{2}\left[(\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\bar\alpha_{t-1}})\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0}{1-\bar\alpha_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} && (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})}\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0}{1-\bar\alpha_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} && (78)\\
&= \exp\left\{-\frac{1}{2}\left[\frac{\alpha_t-\bar\alpha_t+1-\alpha_t}{(1-\alpha_t)(1-\bar\alpha_{t-1})}\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0}{1-\bar\alpha_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} && (79)\\
&= \exp\left\{-\frac{1}{2}\left[\frac{1-\bar\alpha_t}{(1-\alpha_t)(1-\bar\alpha_{t-1})}\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0}{1-\bar\alpha_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\} && (80)\\
&= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar\alpha_t}{(1-\alpha_t)(1-\bar\alpha_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2-2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0}{1-\bar\alpha_{t-1}}\right)}{\frac{1-\bar\alpha_t}{(1-\alpha_t)(1-\bar\alpha_{t-1})}}\boldsymbol{x}_{t-1}\right]\right\} && (81)\\
&= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar\alpha_t}{(1-\alpha_t)(1-\bar\alpha_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2-2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar\alpha_{t-1}}\boldsymbol{x}_0}{1-\bar\alpha_{t-1}}\right)(1-\alpha_t)(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\boldsymbol{x}_{t-1}\right]\right\} && (82)\\
&= \exp\left\{-\frac{1}{2}\left(\frac{1}{\frac{(1-\alpha_t)(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}}\right)\left[\boldsymbol{x}_{t-1}^2-2\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar\alpha_t}\boldsymbol{x}_{t-1}\right]\right\} && (83)\\
&\propto \mathcal{N}(\boldsymbol{x}_{t-1};\underbrace{\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar\alpha_t}}_{\mu_q(\boldsymbol{x}_t,\boldsymbol{x}_0)},\underbrace{\frac{(1-\alpha_t)(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)}) && (84)
\end{aligned}
$$

可以推断，对于每一步的x，都是正态分布的；又，α已知，所以可以把方差改写成式(85)

$$\sigma_q^2(t)=\frac{(1-\alpha_t)(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}$$

又两个高斯分布之间的KL散度为：

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)\parallel\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y))=\frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|}-d+\mathrm{tr}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x)+(\boldsymbol{\mu}_y-\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\mu}_y-\boldsymbol{\mu}_x)\right] \quad (86)$$

由此可得

$$
\begin{aligned}
&\arg\min_{\boldsymbol\theta} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol\theta}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\\
&=\arg\min_{\boldsymbol\theta} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q(t))\parallel\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_{\boldsymbol\theta},\boldsymbol{\Sigma}_q(t))) && (87)\\
&=\arg\min_{\boldsymbol\theta}\frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_q(t)|}{|\boldsymbol{\Sigma}_q(t)|}-d+\mathrm{tr}(\boldsymbol{\Sigma}_q(t)^{-1}\boldsymbol{\Sigma}_q(t))+(\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q)\right] && (88)\\
&=\arg\min_{\boldsymbol\theta}\frac{1}{2}\left[\log 1-d+d+(\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q)\right] && (89)\\
&=\arg\min_{\boldsymbol\theta}\frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q)\right] && (90)\\
&=\arg\min_{\boldsymbol\theta}\frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q)^T\left(\sigma_q^2(t)\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q)\right] && (91)\\
&=\arg\min_{\boldsymbol\theta}\frac{1}{2\sigma_q^2(t)}\left[\|\boldsymbol{\mu}_{\boldsymbol\theta}-\boldsymbol{\mu}_q\|_2^2\right] && (92)
\end{aligned}
$$

可以得到的是，这个式子其实希望优化的是一个与$\mu_q(x_t, x_0)$相符的$\mu_\theta(x_t, t)$，两者分别为

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0)=\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar\alpha_t}$$

$$\boldsymbol{\mu}_{\boldsymbol\theta}(\boldsymbol{x}_t,t)=\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol\theta}(\boldsymbol{x}_t,t)}{1-\bar\alpha_t}$$

其中的ˆxθ(xt, t)是由网络参数化得到的，所以，优化问题就可以简化

$$
\begin{aligned}
&\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\
&= \arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q(t)) \parallel \mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\Sigma}_q(t))) && (95)\\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)}{1-\bar\alpha_t}-\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar\alpha_t}\right\|_2^2\right] && (96)\\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)}{1-\bar\alpha_t}-\frac{\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar\alpha_t}\right\|_2^2\right] && (97)\\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)}{1-\bar\alpha_t}(\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0)\right\|_2^2\right] && (98)\\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\frac{\bar\alpha_{t-1}(1-\alpha_t)^2}{(1-\bar\alpha_t)^2}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (99)
\end{aligned}
$$

总结

综上，优化VDM可以视作学习一个网络，从任意噪声中预测原始的ground truth，且在所有隐藏层上的噪声都要ELBO总和最小

$$\arg\min_{\boldsymbol{\theta}} \mathbb{E}_{t\sim U\{2,T\}}\left[\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]\right]$$

Learning Diffusion Noise Parameters

如果用具有η参数的网络$\alpha_\eta(t)$来学习α，效率很低。解决办法是式(85)代入式(99)化简

$$
\begin{aligned}
\frac{1}{2\sigma_q^2(t)}\frac{\bar\alpha_{t-1}(1-\alpha_t)^2}{(1-\bar\alpha_t)^2}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] &= \frac{1}{2\frac{(1-\alpha_t)(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}}\frac{\bar\alpha_{t-1}(1-\alpha_t)^2}{(1-\bar\alpha_t)^2}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (101)\\
&= \frac{1}{2}\frac{1-\bar\alpha_t}{(1-\alpha_t)(1-\bar\alpha_{t-1})}\frac{\bar\alpha_{t-1}(1-\alpha_t)^2}{(1-\bar\alpha_t)^2}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (102)\\
&= \frac{1}{2}\frac{\bar\alpha_{t-1}(1-\alpha_t)}{(1-\bar\alpha_{t-1})(1-\bar\alpha_t)}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (103)\\
&= \frac{1}{2}\frac{\bar\alpha_{t-1}-\bar\alpha_t}{(1-\bar\alpha_{t-1})(1-\bar\alpha_t)}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (104)\\
&= \frac{1}{2}\frac{\bar\alpha_{t-1}-\bar\alpha_{t-1}\bar\alpha_t+\bar\alpha_{t-1}\bar\alpha_t-\bar\alpha_t}{(1-\bar\alpha_{t-1})(1-\bar\alpha_t)}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (105)\\
&= \frac{1}{2}\frac{\bar\alpha_{t-1}(1-\bar\alpha_t)-\bar\alpha_t(1-\bar\alpha_{t-1})}{(1-\bar\alpha_{t-1})(1-\bar\alpha_t)}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (106)\\
&= \frac{1}{2}\left(\frac{\bar\alpha_{t-1}(1-\bar\alpha_t)}{(1-\bar\alpha_{t-1})(1-\bar\alpha_t)}-\frac{\bar\alpha_t(1-\bar\alpha_{t-1})}{(1-\bar\alpha_{t-1})(1-\bar\alpha_t)}\right)\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (107)\\
&= \frac{1}{2}\left(\frac{\bar\alpha_{t-1}}{1-\bar\alpha_{t-1}}-\frac{\bar\alpha_t}{1-\bar\alpha_t}\right)\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] && (108)
\end{aligned}
$$

根据SNR的定义，$\mathrm{SNR}=\mu^2/\sigma^2$，有

$$\mathrm{SNR}(t)=\frac{\bar\alpha_t}{1-\bar\alpha_t}$$

所以，

$$\frac{1}{2\sigma_q^2(t)}\frac{\bar\alpha_{t-1}(1-\alpha_t)^2}{(1-\bar\alpha_t)^2}\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right]=\frac{1}{2}(\mathrm{SNR}(t-1)-\mathrm{SNR}(t))\left[\left\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\right\|_2^2\right] \qquad (110)$$

信噪比表示原始信号与当前噪声量之间的比值：越高越好。而在diffusion model中，会要求信噪比随着t增加而减少，这样随着时间推移，信号会越来越嘈杂

不妨可以定义信噪比如下

$$\mathrm{SNR}(t)=\exp(-\omega_{\boldsymbol{\eta}}(t))$$

其中w_η是一个依赖于参数η的单调递增函数。由此，可以得到α_t的形式

$$
\begin{aligned}
&\frac{\bar\alpha_t}{1-\bar\alpha_t}=\exp(-\omega_{\boldsymbol{\eta}}(t))\\
&\therefore \bar\alpha_t=\mathrm{sigmoid}(-\omega_{\boldsymbol{\eta}}(t))\\
&\therefore 1-\bar\alpha_t=\mathrm{sigmoid}(\omega_{\boldsymbol{\eta}}(t))
\end{aligned}
$$

三种等效的解释

变分模型可以通过学习网络来训练，但x_0有另外两种等效解释

1.等价于学习预测噪声

利用重参数化技巧

$$\boldsymbol{x}_0=\frac{\boldsymbol{x}_t-\sqrt{1-\bar\alpha_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar\alpha_t}}$$

代入到之前的μ_q，有

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0)=\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar\alpha_t} \tag{116}$$

$$=\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\frac{\boldsymbol{x}_t-\sqrt{1-\bar\alpha_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar\alpha_t}}}{1-\bar\alpha_t} \tag{117}$$

$$=\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+(1-\alpha_t)\frac{\boldsymbol{x}_t-\sqrt{1-\bar\alpha_t}\boldsymbol{\epsilon}_0}{\sqrt{\alpha_t}}}{1-\bar\alpha_t} \tag{118}$$

$$=\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t}{1-\bar\alpha_t}+\frac{(1-\alpha_t)\boldsymbol{x}_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}-\frac{(1-\alpha_t)\sqrt{1-\bar\alpha_t}\boldsymbol{\epsilon}_0}{(1-\bar\alpha_t)\sqrt{\alpha_t}} \tag{119}$$

$$=\left(\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}+\frac{1-\alpha_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\right)\boldsymbol{x}_t-\frac{(1-\alpha_t)\sqrt{1-\bar\alpha_t}}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \tag{120}$$

$$=\left(\frac{\alpha_t(1-\bar\alpha_{t-1})}{(1-\bar\alpha_t)\sqrt{\alpha_t}}+\frac{1-\alpha_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\right)\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \tag{121}$$

$$=\frac{\alpha_t-\bar\alpha_t+1-\alpha_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \tag{122}$$

$$=\frac{1-\bar\alpha_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \tag{123}$$

$$=\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0 \tag{124}$$

由此，μ_θ与θ有关，可以设

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)=\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$$

所以

$$\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))$$

$$=\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q(t))\parallel\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\Sigma}_q(t))) \tag{126}$$

$$=\arg\min_{\boldsymbol{\theta}}\frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0\right\|_2^2\right] \tag{127}$$

$$=\arg\min_{\boldsymbol{\theta}}\frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}_0-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\right\|_2^2\right] \tag{128}$$

$$=\arg\min_{\boldsymbol{\theta}}\frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}\sqrt{\alpha_t}}(\boldsymbol{\epsilon}_0-\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t))\right\|_2^2\right] \tag{129}$$

$$=\arg\min_{\boldsymbol{\theta}}\frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{(1-\bar\alpha_t)\alpha_t}\left[\left\|\boldsymbol{\epsilon}_0-\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\right\|_2^2\right] \tag{130}$$

此处的ε_θ预测高斯状的源噪声ε_0，这决定了x_0如何到x_t。因此，这个式子说明通过预测原始图像x_0来学习VDM等价于学习预测噪声，甚至性能更好；

2.等价于建模源噪声的负数

利用Tweedie's Formula，一个给定样本的指数族分布的真实均值可以通过样本的最大似然估计（经验平均值）加上修正项得到

这可以用来减轻样本偏差

Tweedie's Formula

对于一个高斯型的变量z有如下表述

$$\mathbb{E}\left[\boldsymbol{\mu}_z|\boldsymbol{z}\right]=\boldsymbol{z}+\boldsymbol{\Sigma}_z\nabla_{\boldsymbol{z}}\log p(\boldsymbol{z})$$

对于目前的问题，显然

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

所以，利用Tweedie's Formula可以得到

$$\mathbb{E}\left[\boldsymbol{\mu}_{x_t}|\boldsymbol{x}_t\right]=\boldsymbol{x}_t+(1-\bar\alpha_t)\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)$$

可以估计生成的x_t的真实均值，所以两式联立有

$$\sqrt{\bar\alpha_t}\boldsymbol{x}_0=\boldsymbol{x}_t+(1-\bar\alpha_t)\nabla\log p(\boldsymbol{x}_t)$$

$$\therefore\boldsymbol{x}_0=\frac{\boldsymbol{x}_t+(1-\bar\alpha_t)\nabla\log p(\boldsymbol{x}_t)}{\sqrt{\bar\alpha_t}}$$

由此，可以把x_0代入到ground truth去噪过渡均值μ_q

$$\begin{aligned}
\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0) &= \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar\alpha_t} && (134)\\
&= \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\frac{\boldsymbol{x}_t+(1-\bar\alpha_t)\nabla\log p(\boldsymbol{x}_t)}{\sqrt{\bar\alpha_t}}}{1-\bar\alpha_t} && (135)\\
&= \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+(1-\alpha_t)\frac{\boldsymbol{x}_t+(1-\bar\alpha_t)\nabla\log p(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}}{1-\bar\alpha_t} && (136)\\
&= \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t}{1-\bar\alpha_t}+\frac{(1-\alpha_t)\boldsymbol{x}_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}+\frac{(1-\alpha_t)(1-\bar\alpha_t)\nabla\log p(\boldsymbol{x}_t)}{(1-\bar\alpha_t)\sqrt{\alpha_t}} && (137)\\
&= \left(\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}+\frac{1-\alpha_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\right)\boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log p(\boldsymbol{x}_t) && (138)\\
&= \left(\frac{\alpha_t(1-\bar\alpha_{t-1})}{(1-\bar\alpha_t)\sqrt{\alpha_t}}+\frac{1-\alpha_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\right)\boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log p(\boldsymbol{x}_t) && (139)\\
&= \frac{\alpha_t-\bar\alpha_t+1-\alpha_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log p(\boldsymbol{x}_t) && (140)\\
&= \frac{1-\bar\alpha_t}{(1-\bar\alpha_t)\sqrt{\alpha_t}}\boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log p(\boldsymbol{x}_t) && (141)\\
&= \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log p(\boldsymbol{x}_t) && (142)
\end{aligned}$$

因此，μ_θ类似地也有

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)=\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{\alpha_t}}\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$$

所以，对应地优化问题就变成了

$$\begin{aligned}
&\underset{\boldsymbol{\theta}}{\arg\min}\, D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\\
&=\underset{\boldsymbol{\theta}}{\arg\min}\, D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q(t))\parallel\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\Sigma}_q(t)))\\
&=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{\alpha_t}}\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log p(\boldsymbol{x}_t)\right\|_2^2\right]\\
&=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1-\alpha_t}{\sqrt{\alpha_t}}\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla\log p(\boldsymbol{x}_t)\right\|_2^2\right]\\
&=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2\sigma_q^2(t)}\left[\left\|\frac{1-\alpha_t}{\sqrt{\alpha_t}}(\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\nabla\log p(\boldsymbol{x}_t))\right\|_2^2\right]\\
&=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{\alpha_t}\left[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)-\nabla\log p(\boldsymbol{x}_t)\right\|_2^2\right]
\end{aligned}$$

其中，s_θ是一个神经网络，学习预测评分函数$\nabla_x \log p(x_t)$，这是空间中对x_t的梯度。且这个评分函数在形式上与ϵ_0很类似：

$$\begin{aligned}
\boldsymbol{x}_0=\frac{\boldsymbol{x}_t+(1-\bar\alpha_t)\nabla\log p(\boldsymbol{x}_t)}{\sqrt{\bar\alpha_t}}&=\frac{\boldsymbol{x}_t-\sqrt{1-\bar\alpha_t}\boldsymbol{\epsilon}_0}{\sqrt{\bar\alpha_t}}\\
\therefore (1-\bar\alpha_t)\nabla\log p(\boldsymbol{x}_t)&=-\sqrt{1-\bar\alpha_t}\boldsymbol{\epsilon}_0\\
\nabla\log p(\boldsymbol{x}_t)&=-\frac{1}{\sqrt{1-\bar\alpha_t}}\boldsymbol{\epsilon}_0
\end{aligned}$$

这两项只相差一个随时间变化的常数因子。这说明，评分函数是度量如何在数据空间中移动来最大化logp(x)的。源噪声添加到图像中来破坏它，那么往相反方向移动就是对图像去噪。所以，学习建模评分函数等价于建模源噪声的负数；

总结

由此就有3种对于优化VDM的不同解释：1.学习网络来预测原始图像x_0、预测源噪声ϵ_0、或者是预测任意t下的评分函数$\nabla \log p(x_t)$。

基于分数的生成模型

在上文的第三段有说明优化VDM就是预测评分函数。但是，这并不能直观地体现出评分函数到底是什么。为此，有必要介绍一下基于分数的生成模型；

为了理解优化评分函数的意义，需要先重新审视这个基于能量的模型。对于任意的概率分布，有

$$p_{\boldsymbol{\theta}}(\boldsymbol{x})=\frac{1}{Z_{\boldsymbol{\theta}}}e^{-f_{\boldsymbol{\theta}}(\boldsymbol{x})}$$

其中，$f_\theta(x)$是一个带参数的能量函数，由网络建模的得到；Z_θ是归一化常数，确保概率不超过1；这种形式显然无法得到分布的信息，解决办法是用神经网络学习评分函数$\nabla \log p(x)$，而不是p(x)。对式(152)求导可以得到：

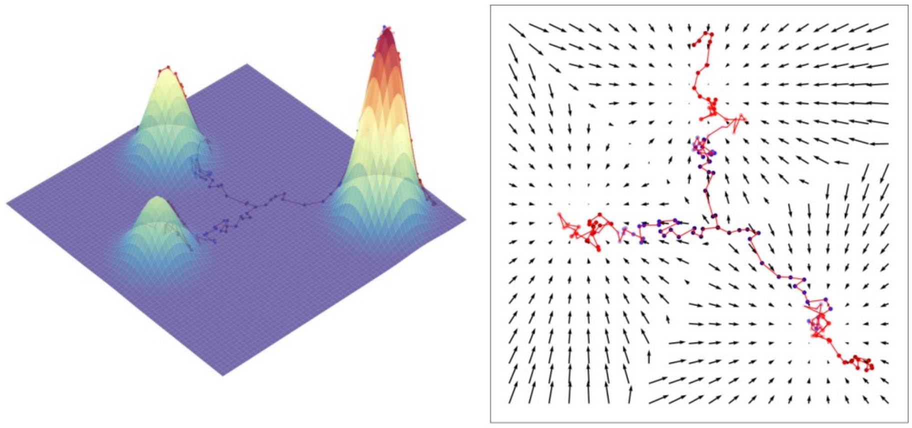
$$
\begin{aligned}
\nabla_{\boldsymbol{x}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \nabla_{\boldsymbol{x}} \log\left(\frac{1}{Z_{\boldsymbol{\theta}}} e^{-f_{\boldsymbol{\theta}}(\boldsymbol{x})}\right) \\
&= \nabla_{\boldsymbol{x}} \log \frac{1}{Z_{\boldsymbol{\theta}}} + \nabla_{\boldsymbol{x}} \log e^{-f_{\boldsymbol{\theta}}(\boldsymbol{x})} \\
&= -\nabla_{\boldsymbol{x}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \\
&\approx \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x})
\end{aligned}
$$

这个函数易于让网络学习，且不涉及归一化常数。而这个评分函数可以如下得到：

$$
\mathbb{E}_{p(\boldsymbol{x})}\left[\left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla \log p(\boldsymbol{x})\right\|_2^2\right]
$$

那么，这个评分函数代表了什么？评分函数对每个x取了梯度的对数似然，这本质上描述了它的似然在数据空间中移动进而增加的方向；

直观上来看，评分函数在x的空间上定义了一个指向模态modes的梯度场，见下图右侧



通过学习真实数据的分布，可以从空间中的任意一点，沿着评分函数生成样本，直到达到同一种模式，这就是Langevin动力学：

$$
\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + c\nabla \log p(\boldsymbol{x}_i) + \sqrt{2c}\boldsymbol{\epsilon}, \quad i = 0, 1, ..., K
$$

其中x0从先验分布如均匀分布中随机采样，加上噪声保证不总会收敛到同一个模式；又评分函数是确定的，所以抽样噪声可以增加随机性；

注意，对于比如图像这种复杂分布，要得到ground truth的评分函数显然不可行的；可以用分数匹配score matching的方式，通过最小化Fisher散度来得到。

下略

Guidance

guidance其实就是添加了条件信息。一种思路是在每次迭代时和timestep并排添加，考虑下面这个式子：

$$
p(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)
$$

为了变成条件信息，加入y：

$$
p(\boldsymbol{x}_{0:T}|y) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, y)
$$

比如，y可以是图像-文本生成中的文本编码；但这样可能会在迭代中学会忽略条件信息，需要改进。

Classifier Guidance

利用评分函数，引入y，有

$$
\begin{aligned}
\nabla \log p(\boldsymbol{x}_t|y) &= \nabla \log\left(\frac{p(\boldsymbol{x}_t)p(y|\boldsymbol{x}_t)}{p(y)}\right) \\
&= \nabla \log p(\boldsymbol{x}_t) + \nabla \log p(y|\boldsymbol{x}_t) - \nabla \log p(y) \\
&= \underbrace{\nabla \log p(\boldsymbol{x}_t)}_{\text{unconditional score}} + \underbrace{\nabla \log p(y|\boldsymbol{x}_t)}_{\text{adversarial gradient}}
\end{aligned}
$$

一部分是没有条件的评分函数，一部分是classifier。用一个超参数γ来控制，可以得到：

$$
\nabla \log p(\boldsymbol{x}_t|y) = \nabla \log p(\boldsymbol{x}_t) + \gamma \nabla \log p(y|\boldsymbol{x}_t)
$$

当γ=0，相当于忽略条件信息；当γ很大，说明很依赖于条件信息。

缺点是需要处理单独学习分类器。

Classifier-Free Guidance

对于Classifier Guidance进行重排，有

$$\nabla \log p(y|\boldsymbol{x}_t) = \nabla \log p(\boldsymbol{x}_t|y) - \nabla \log p(\boldsymbol{x}_t)$$

代入γ控制的guidance，有

$$\begin{aligned}\nabla \log p(\boldsymbol{x}_t|y) &= \nabla \log p(\boldsymbol{x}_t) + \gamma \left(\nabla \log p(\boldsymbol{x}_t|y) - \nabla \log p(\boldsymbol{x}_t)\right) \\ &= \nabla \log p(\boldsymbol{x}_t) + \gamma \nabla \log p(\boldsymbol{x}_t|y) - \gamma \nabla \log p(\boldsymbol{x}_t) \\ &= \underbrace{\gamma \nabla \log p(\boldsymbol{x}_t|y)}_{\text{conditional score}} + \underbrace{(1-\gamma) \nabla \log p(\boldsymbol{x}_t)}_{\text{unconditional score}}\end{aligned}$$

此时γ就是控制一个条件模型对条件信息关系程度的项。当γ=0，相当于忽略条件信息；当γ很大，说明很依赖于条件信息。

总结

1.变分扩散模型VDM是马尔可夫分层变分自动编码器MVHA的特例

2.利用三个假设让ELBO可以计算

3.分析了优化VDM的三种角度

4.引入先验就是引入条件信息分布