Guest Lecture 23: Deep Generative Models: Diffusion Basics

Junming Cao
SIST, ShanghaiTech
Fall, 2022

Basic Diffusion Model

一、条件概率公式与高斯分布的KL散度

1. 条件概率的一般形式

$P(A,B,C)=P(C|B,A)P(B,A)=P(C|B,A)P(B|A)P(A)$

$P(B,C|A)=P(A,B,C)/P(A)=P(B|A)P(C|A,B)$

$P(A|B)=P(B|A)\cdot\frac{P(A)}{P(B)}$

2. 基于马尔科夫假设的条件概率

如果满足马尔科夫链关系A→B→C，则有

$P(A,B,C)=P(C|B,A)P(B|A)P(A)=P(C|B)P(B|A)P(A)$

$P(B,C|A)=P(B|A)P(C|B)$

3. 高斯分布的KL散度公式

KL散度：

对于两个单一变量的高斯分布P和Q而言，它们的KL散度为

$$\mathrm{KL}(P,Q)=log\frac{\sigma_2}{\sigma_1}+\frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}-\frac{1}{2}$$

KL-Divergence

4. 参数重整化

若希望从高斯分布$N(\mu,\sigma)$中采样，可以先从标准分布$N(0,1)$中采样出z，再得到$\sigma\times z+\mu$。
这样做的好处是将随机性转移到了z这个常量上，而μ与σ则当做仿射变换网络的一部分

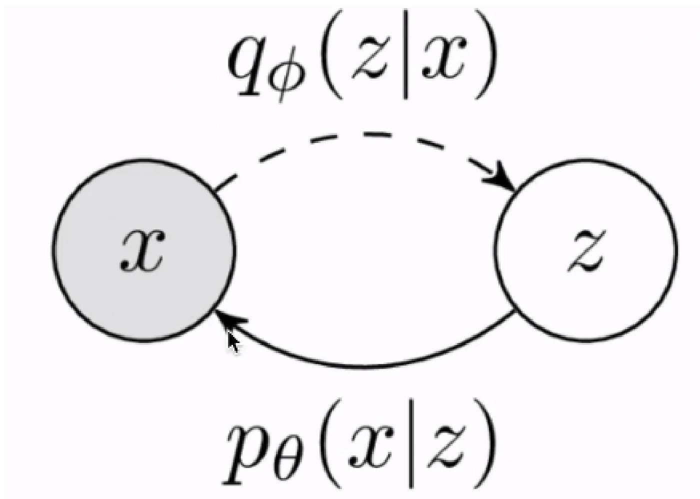
在VAE和Diffusion中大量运用

二、VAE与多层VAE回顾

0. AE(Auto Encoder)自编码器回顾

Auto Encoder

1. 单层VAE的原理公式与置信下界



训练时通过X生成Z，$Z=q_\phi(X)$，$q_\phi(z|x)$为概率编码器
推理时通过Z预测X，$X=p_\theta(Z)$，$p_\theta(x|z)$为概率解码器

联合概率分布对z进行积分得到边缘分布：$p_\theta(x)=\int_z p_\theta(x,z)=\int_z p_\theta(x|z)p_\theta(z)dz$。

对联合概率分布上下同成后验概率分布：

$$\int_z q_\phi(z|x)\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}dz$$

即$\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}$在$q_\phi(z|x)$下的期望，再两边取log：

$$\log(p_\theta(x))=\log(\mathbb{E}_{z\sim q_\phi(z|x)}[\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}])$$

根据Jensen不等式：

$$\log(p(x))\geq\mathbb{E}_{z\sim q_\phi(z|x)}[\log\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}]$$

$$=\mathbb{E}_{z\sim q_\phi(z|x)}[\log(p_\theta(x|z)]+\mathbb{E}_{z\sim q_\phi(z|x)}[\log\frac{p_\theta(z)}{q_\phi(z|x)}]$$

$$=\mathbb{E}_{z\sim q_\phi(z|x)}[\log(p_\theta(x|z)]-\mathrm{D}_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z))$$

右侧即为置信下界。

第一项为 reconstruction term，重构项

第二项为 prior matching term

训练目标为最大化$\log(p(x))$，最大化下界即可最大化$\log(p(x))$

另一种置信下界推导方式（二者等价）

展开反向KL散度公式：

$$
\begin{aligned}
&\mathrm{D_{KL}}(q_\phi(z|x)||p_\theta(z|x)) \\
&= \int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(z|x)}dz \\
&= \int q_\phi(z|x)\log\frac{q_\phi(z|x)p_\theta(x)}{p_\theta(z,x)}dz \\
&= \int q_\phi(z|x)[\log p_\theta(x) + \log\frac{q_\phi(z|x)}{p_\theta(z,x)}]dz \\
&= \log p_\theta(x) + \int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(x|z)p_\theta(z)}dz \qquad \text{Because } \int q(z|x)q(z)dz = 1 \\
&= \log p_\theta(x) + \int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(z)}dz - \int q_\phi(z|x)\log(p_\theta(x|z)dz \\
&= \log p_\theta(x) + \mathrm{D_{KL}}(q_\phi(z|x)||p_\theta(z)) - \mathbb{E}_{z\sim q_\phi(z|x)}\log(p_\theta(x|z))
\end{aligned}
$$

重新排列方程左右：

$$\log p_\theta(x) - \mathrm{D_{KL}}(q_\phi(z|x)||p_\theta(z|x) = \mathbb{E}_{z\sim q_\phi(z|x)}\log(p_\theta(x|z)) - \mathrm{D_{KL}}(q_\theta(z|x)||p_\theta(z))$$

等号左侧是学习真实分布时想最大化的东西：产生真实数据的可能性$p_\theta(x)$，同时最小化真实分布和后验分布（$q_\phi(z|x)$）之间的差距。相对于q_ϕ，$p_\theta(x)$是固定的。

同时等号左侧的负值即为损失函数。

$$\mathrm{L_{VAE}}(\theta,\phi) = -\log p_\theta(x|z) + \mathrm{D_{KL}}(q_\phi(z|x)||p_\theta(z|x))$$

$$= -\mathbb{E}_{z\sim q_\phi(z|x)}\log p_\theta(x|z) + \mathrm{D_{KL}}(q_\phi(z|x)||p_\theta(z))$$

$$(\theta^*,\phi^*) = \arg\min_{\theta,\phi}\mathrm{L_{VAE}}$$

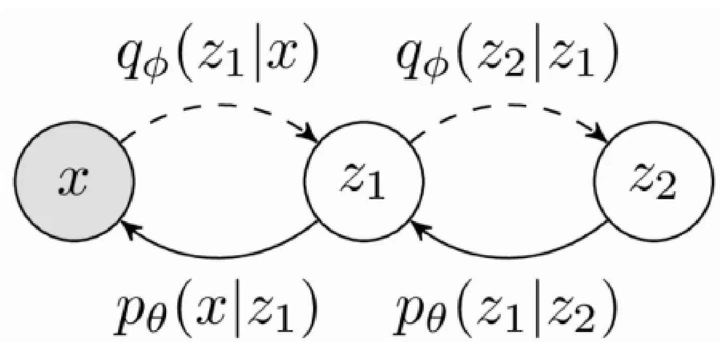
在变分贝叶斯方法中，这个损失函数被称为变分下界。$-\mathrm{L_{VAE}}$就是$\log p_\theta(x)$的下界。

$$-\mathrm{L_{VAE}} = \log p_\theta(x) - \mathrm{D_{KL}}(q_\theta(z|x)p_\theta(z|x) \leq \log p_\theta(x)$$

即通过最小化损失可以最大限度地提升生成真实数据样本的概率下界。

2. 多层VAE的原理公式与置信下界

2.1 双层VAE：



$$p_\theta(x) = \iint_{z_1,z_2} p_\theta(x, z_1, z_2)dz_1, dz_2$$

$$p_\theta(x) = \iint q_\phi(z_1, z_2|x)\frac{p_\theta(x, z_1, z_2)}{q_\phi(z_1, z_2|x)}dz_1 dz_2$$

$$\log(p_\theta(x)) = \mathbb{E}_{z_1,z_2\sim q_\phi(z_1,z_2|x)}[\log \frac{p_\theta(x, z_1, z_2)}{q_\phi(z_1, z_2|x)}]$$

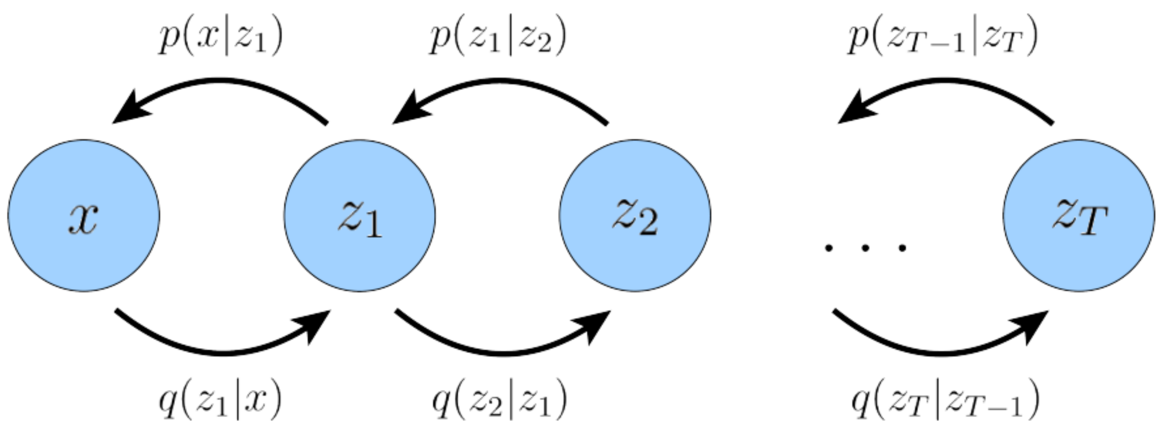
利用Jensen不等式

$$\log p(x) \geq \mathbb{E}_{z_1,z_2\sim q_\phi(z_1,z_2|x)}[\log \frac{p_\theta(x, z_1, z_2)}{q_\phi(z_1, z_2|x)}]$$

利用马尔科夫链

$$\mathrm{L}(\theta, \phi) = \mathbb{E}_{z_1,z_2\sim q_\phi(z_1,z_2|x)}[\log p_\theta(x|z_1) + \log p_\theta(z_1|z_2) + \log p_\theta(z_2) - \log q_\phi(z_2|z_1) - \log q_\phi(z_1|x)]$$

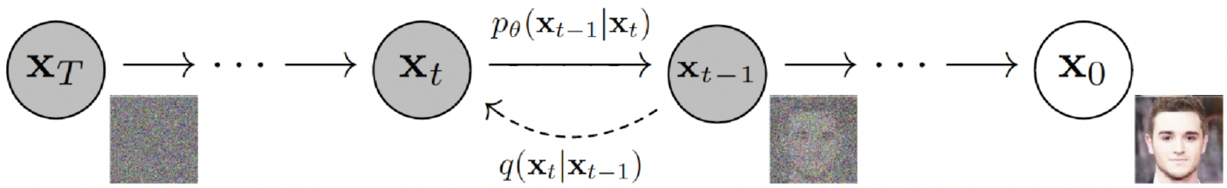
2.2 多层VAE



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

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

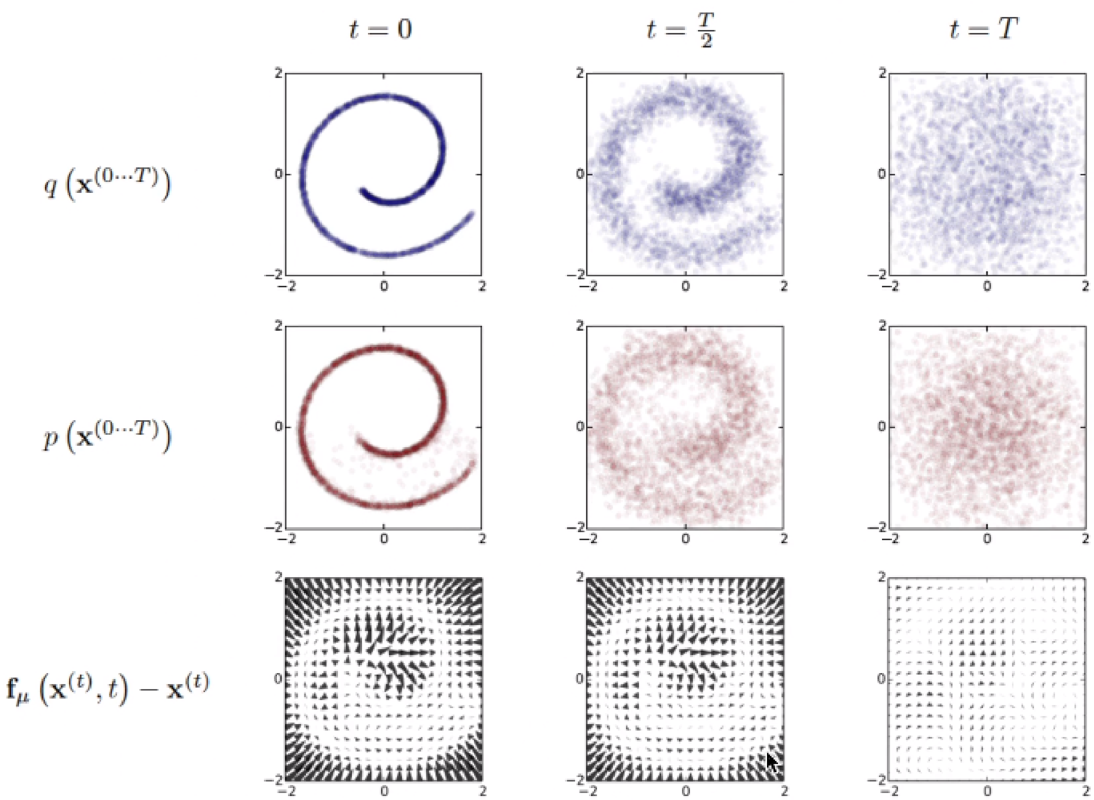
三、Diffusion Model图示



当满足以下三个条件时，可以将Variational Diffusion Models视作马尔科夫条件下的VAE：

1. latent层的维度和数据维度完全一致；
2. 每个t的latent encoder将不作为一个可学习变量，而是严格的线性高斯模型；
3. 最终T时刻的latent是标准正态分布。

与多层VAE类似，层层概率推导，有理由相信最终的cost function形式将类似**多层VAE**的cost function



扩散过程：从x_0逐渐到X_T的熵增过程，最终为各向异的高斯分布，训练（正向）过程
逆扩散过程：反向的过程，推理过程
漂移量：$\mathrm{f}_\mu(\mathrm{x}^{(t)}, t) - \mathrm{x}^{(t)}$，推理和训练过程的状态差

四、扩散过程（Diffusion Process）

1. 给定初始数据分布$x_0 \sim q(x)$，不断向分布中添加高斯噪声（仿射变换）。
   噪声方差：$\beta_t \in [0, 1]$

均值：固定值β_t和当前时刻t的数据x_t共同决定
方差和均值都是确定的，不含参，为超参
马尔科夫过程

$$Z \sim \mathcal{N}(0,1) \quad X_t = Z \times \sqrt{\beta_t} + \sqrt{1-\beta_t} X_{t-1}$$

2. 随着t不断增大，最终数据分布x_T变为各向独立的高斯分布

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}) \quad q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

3. 任意时刻的$q(x_t)$推导也可以完全基于x_0和β_t计算出来，不需要做迭代
设$\alpha_t = 1 - \beta_t$，$\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$
同时存在定理：若X, Y互相独立且都属于高斯分布，$X \sim \mathcal{N}(\mu_1, \sigma_1)$，$Y \sim \mathcal{N}(\mu_2, \sigma_2)$，则$aX + bY \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\mu_1^2 + b^2\mu_2^2)$。证明 Proof of distribution of aX+bY
则：$X_t = \sigma Z_{t-1} + \mu$

$$\begin{aligned}
&= \sqrt{1-\alpha_t} Z_{t-1} + \sqrt{\alpha_t} X_{t-1}; \quad Z_{t-1}, Z_{t-2}, ... \sim \mathcal{N}(0, \mathbf{I}) \\
&= \sqrt{1-\alpha_t} Z_{t-1} + \sqrt{\alpha_t} \cdot [\sqrt{1-\alpha_{t-1}} Z_{t-2} + \sqrt{\alpha_{t-1}} X_{t-2}] \\
&= \sqrt{1-\alpha_t} Z_{t-1} + \sqrt{\alpha_t(1-\alpha_{t-1})} Z_{t-2} + \sqrt{\alpha_t \alpha_{t-1}} X_{t-2} \\
&\qquad \sigma_1 = \sqrt{1-\alpha_t}, \quad \sigma_2 = \sqrt{\alpha_t(1-\alpha_{t-1})} \\
&\bar{\sigma} = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{1-\alpha_t + \alpha_t - \alpha_t\alpha_{t-1}} = \sqrt{1-\alpha_t\alpha_{t-1}} \\
&= \bar{\sigma} \bar{Z}_{t-2} + \sqrt{\alpha_t \alpha_{t-1}} X_{t-2} \\
&= \sqrt{1-\alpha_t\alpha_{t-1}} \bar{Z}_{t-2} + \sqrt{\alpha_t\alpha_{t-1}} X_{t-2} \\
&= \sqrt{1-\alpha_t\alpha_{t-1}} \bar{Z}_{t-2} + \sqrt{\alpha_t\alpha_{t-1}} \cdot [\sqrt{\alpha_{t-2}} X_{t-3} + \sqrt{1-\alpha_{t-2}} Z_{t-2}] \\
&= \sqrt{1-\alpha_t\alpha_{t-1}\alpha_{t-2}} \bar{Z}_{t-3} + \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}} X_{t-3} \\
&= \sqrt{1-\bar{\alpha}_t} \bar{Z} + \sqrt{\bar{\alpha}_t} X_0
\end{aligned}$$

4. β_t的取值策略：样本（X_t）中的噪声越多，β_t越大，即：

$$\beta_1 < \beta_2 < ... < \beta_T; \quad \bar{\alpha}_1 > \bar{\alpha}_2 > ... > \bar{\alpha}_T$$

五、逆扩散过程（Reverse Process）

从噪声中恢复出原始数据的过程。

由第三个限制条件我们可知最终的latent概率$p(x_T)$是标准正态分布

$$p_\theta(X_{t-1}|X_t) \sim \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \Sigma_\theta(X_t, t))$$

六、后验的扩散条件概率

$$q(X_{1:T}|X_0) = \prod_{t=1}^{T} q(X_t|X_{t-1})$$

$$p(X_{0:T})=p(X_T)\prod_{t=1}^{T}p_\theta(X_{t-1}|X_t)\quad p(X_T)=\mathcal{N}(X_T;\mathbf{0},\mathbf{I})$$

6.1

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T})d\boldsymbol{x}_{1:T}\\
&= \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}\\
&= \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]\\
&\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]\\
&= \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]\\
&= \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]\\
&= \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]\\
&= \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]\\
&= \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]\\
&= \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]\\
&= \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]\\
&= \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{\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}}\\
&\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_{\boldsymbol{\theta}}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}
\end{aligned}$$

1. reconstruction term：预测了逆扩散过程中最后一步到结果的后验概率，训练方法与传统VAE类似；
2. prior matching term: 当最终的latent code满足高斯分布时可以最小化。传统VAE中也存在这一项，但是不同的是，diffusion model中的此项没有可训练项，且最后T时刻一定是各向异性的高斯分布，即$p(x_T)=q(x_T|x_{T-1})=1$，所以此项为0；
3. consistency term: 训练使得 为一张噪音更多的照片去噪的一步 与 从一张噪音更少的照片中添加噪音的过程保持一直。随着训练$p_\theta(x_t|x_{t+1})$吻合$q(x_t|x_{t+1})$，此项的值也在变小。此项为决定项。

6.2

换一个一次一步的算法。

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0)$$

根据贝叶斯法则，改写为：

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

$$\begin{aligned}\log p(\boldsymbol{x}) &\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \\ &= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]\end{aligned}$$

从此步开始推导发生变化：

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

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

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

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

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

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

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

与上一种推导方式作对比：

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\theta}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T))\right]}_{\text{prior matching term}}$$
$$- \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_{\theta}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}$$

$x_t \sim q(x_t|x_{t-1})$可以改写成

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

同样可以推导至

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0$$

$$
\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)}\\
&= \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})}\\
&\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\}\\
&= \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\}\\
&= \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\}\\
&\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\}\\
&= \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\}\\
&= \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\}\\
&= \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\}\\
&= \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\}\\
&= \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\}\\
&= \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\}\\
&= \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\}\\
&\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}}_{\mathbf{\Sigma}_q(t)})
\end{aligned}
$$

证明了每一步$X_{t-1}\sim q(x_{t-1}|x_t,x_0)$均为正态分布，且均值为$x_t$和$x_0$的函数，方差为$\alpha$的函数

令

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

我们要最小化这一项：

$$\sum_{t=2}^{T}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

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

$\mu_q = \mu_q(x_t, x_0),\ \ \mu_\theta = \mu_\theta(x_t, t)$

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

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

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

Summary

- Diffusion Basics
  - From VAE to Diffusion
- Next time
  - Diffusion variants and applications

- Quiz9: send to https://www.gradescope.com/courses/454988/assignments/2502149/submissions
- Keep working on the projects!