扩散生成模型 Diffusion Model理论基础

References:

《Understanding Diffusion Models: A Unified Perspective》
https://arxiv.org/abs/2208.11970
《Denoising Diffusion Probabilistic Models》
https://arxiv.org/abs/2006.11239

目录

- VAE数学原理及模型结构
- MHVAE理论推导
- VDM理论推导(DDPM)

VAE数学原理及模型结构

Variational AutoEncoder

VAE(Variational Autoencoders)简介

Likelihood-based生成模型：给定一个数据集x_D，训练使得模型最大化likelihood $p_\phi(x_D)$

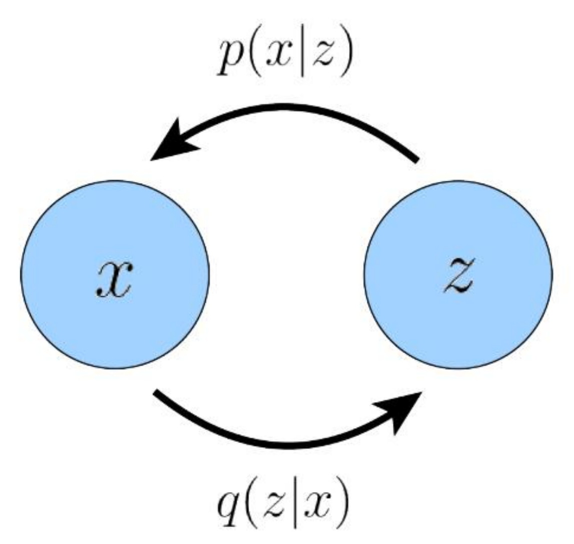
$$x_D \xrightarrow[\phi]{\text{model}} p_\phi(x) \xrightarrow{\text{sample}} x'$$

VAE： 生成需要采样，借助一个变量z和自定义分布$\mathrm{p}(z)$，一般选择多维标准高斯分布 $N\sim(z;\mathbf{0},\boldsymbol{I})$

$\mathrm{q}(\mathrm{z}|x)$：Encoder
$\mathrm{p}(x|\mathrm{z})$：Decoder
将z和x建立起了联系

对高斯分布的z采样，就能通过decoder得到一个新的生成数据 x'

z：latent variable



VAE数学原理

VAE模型优化目标：$p_\phi(x)$ 趋近于真实数据分布 $p_\phi(x) \rightarrow p(x)$

$\log P(x)$有下界：

$$\boxed{\log p(\boldsymbol{x})} \geq \boxed{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right]} \quad (1)$$

ELBO: Evidence Lower BOnd

ELBO公式证明方法一（琴生不等式）：

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x},\boldsymbol{z})d\boldsymbol{z} && \text{边缘化} \\
&= \log \int \frac{p(\boldsymbol{x},\boldsymbol{z})q_\phi(\boldsymbol{z}|\boldsymbol{x})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}d\boldsymbol{z} && \times 1 \\
&= \log \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{期望的定义} \\
&\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{琴生不等式}
\end{aligned}$$

期望的定义：$E_{P(B)}\left[\frac{p(A)}{p(B)}\right] = \int \frac{p(A)}{p(B)} \cdot p(B)\mathrm{d}B$

琴生不等式：https://upload.wikimedia.org/wikipedia/commons/transcoded/5/52/Convex_01.ogv/Convex_01.ogv.360p.webm

VAE数学原理

ELBO是$\log P(x)$下界，但他们之间具体有什么关系呢？
什么时候能取到等号？
为什么说：　优化VAE ⇔ 最大化ELBO

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

ELBO公式证明方法二：

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log p(\boldsymbol{x}) \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) dz \\
&= \int q_\phi(\boldsymbol{z}|\boldsymbol{x})(\log p(\boldsymbol{x})) dz \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x})\right] \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}\right] \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] \\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}|\boldsymbol{x})}\right] \\
&= \boxed{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] + D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x}))} \\
&\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right]
\end{aligned}$$

- ×1,任一分布概率密度函数的积分等于1 ⇐ $\int p(A)dA = 1$
- 期望的定义 ⇐ $E_{P(B)}[p(A)] = \int p(A)\cdot p(B)dB$
- 链式法则 ⇐ $p(x,z) = p(x)\cdot p(z|x)$
- ×1
- 和的期望=期望的和 ⇐ $E[log(AB)] = E[log(A)]+E[log(B)]$
- KL散度定义 ⇐ $D_{KL}[p(A)||p(B)] = E_{P(A)}\left[log\frac{p(A)}{p(B)}\right]$
- KL散度≥0

VAE数学原理

$$\log p(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] + D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}|\boldsymbol{x}))$$

ELBO是$\log P(x)$下界，但他们之间具体有什么关系呢?

答：差1个KL散度，模型encoder拟合的$q_\phi(z|x)$和真实的$p(z|x)$之间的KL散度

什么时候能取到等号?

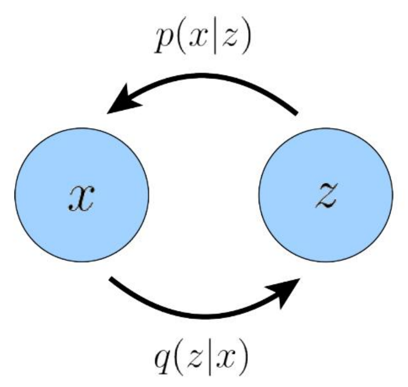
答：KL散度=0，模型encoder完美拟合真实的$p(z|x)$，及$q_\phi(z|x) = p(z|x)$

为什么说： 优化VAE ⇔ 最大化ELBO?

答：VAE的模型框架就是拟合数据变量 x 和隐变量 z 之间的联系。
也就是通过decoder和encoder分别拟合$p(x|z)$和$q(z|x)$，间接拟合$p_\phi(x) \rightarrow p(x)$

KL散度项反映的就是encoder对于$p(z|x)$的拟合程度，我们要最小化它。但是直接优化这个KL散度不可行，因为我们不知道$p(z|x)$这个groundtruth。

$\log p(x)$是真实数据分布的概率，给定一个数据集，该值就唯一确定，与模型无关。
所以：最小化KL散度 ⇔ 最大化ELBO。



VAE数学原理

我们继续拆解ELBO，看看其有什么具体含义？

$$\begin{aligned}\mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\right] &= \mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\right] \\ &= \mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] + \mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z})}{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\right] \\ &= \underbrace{\mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q_{\phi}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))}_{\text{prior matching term}}\end{aligned}$$

链式法则 ← $p(x,z) = p(x) \cdot p(z|x)$

和的期望=期望的和

KL散度定义 ← $D_{KL}[p(A)||p(B)] = E_{P(A)}\left[log\frac{p(A)}{p(B)}\right]$

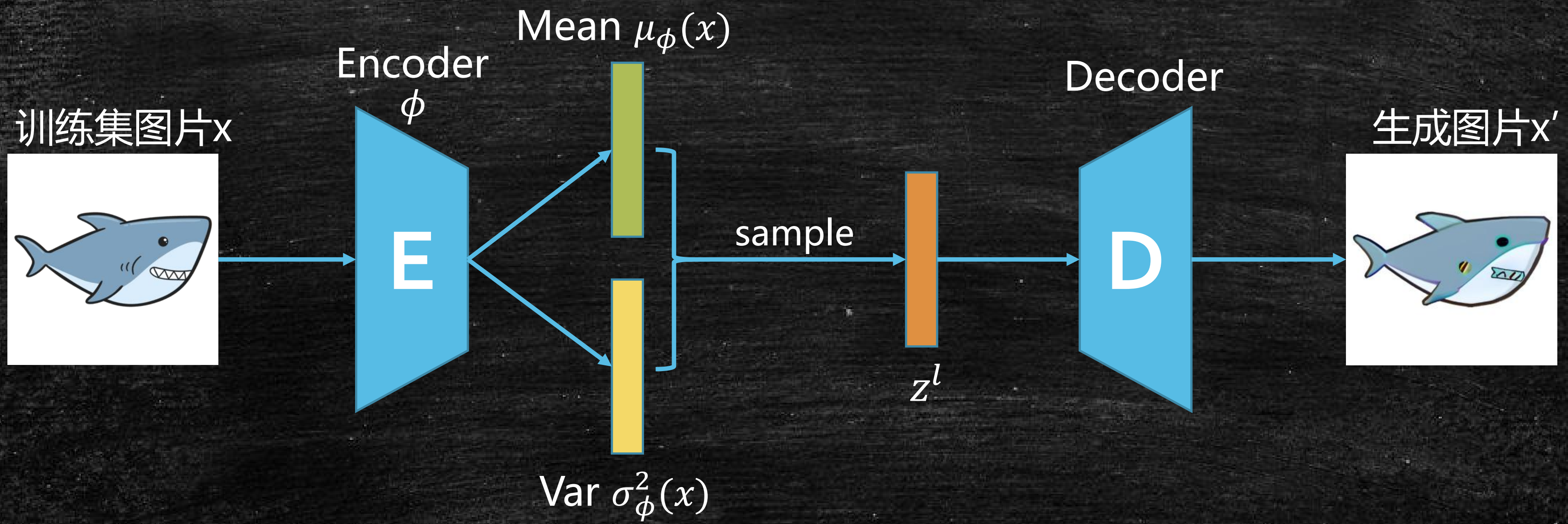
最大化Reconstruction term：让decoder能最大可能从隐变量z生成原始的真实数据x

最小化Prior matching term：让encoder将真实数据x映射到隐变量z后，z尽量满足我们指定的分布，一般为多维标准高斯分布 $N \sim (z; \mathbf{0}, \boldsymbol{I})$

TIPs：如果把Prior matching term这项损失去掉，就是AE模型，AE的z分布未知，没法有效采样，所以也不能作为生成模型

VAE模型结构

VAE模型结构具体是怎么实现的？与ELBO的两部分对应关系是什么？



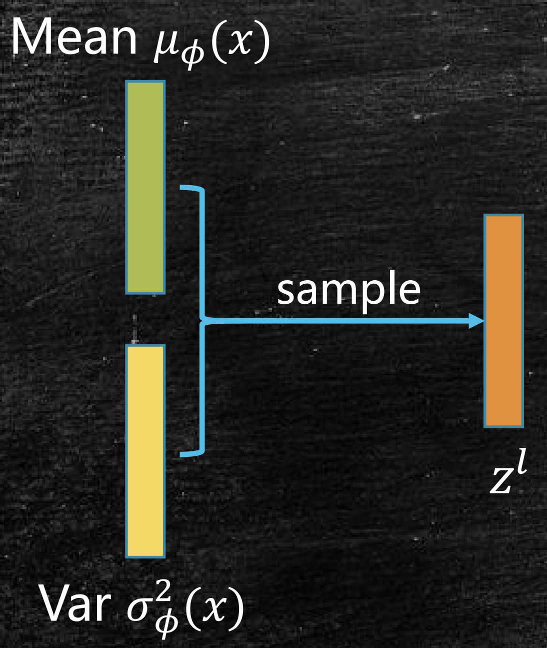
最大化Reconstruction term：生成图片x′ 和训练集图片x尽可能趋近

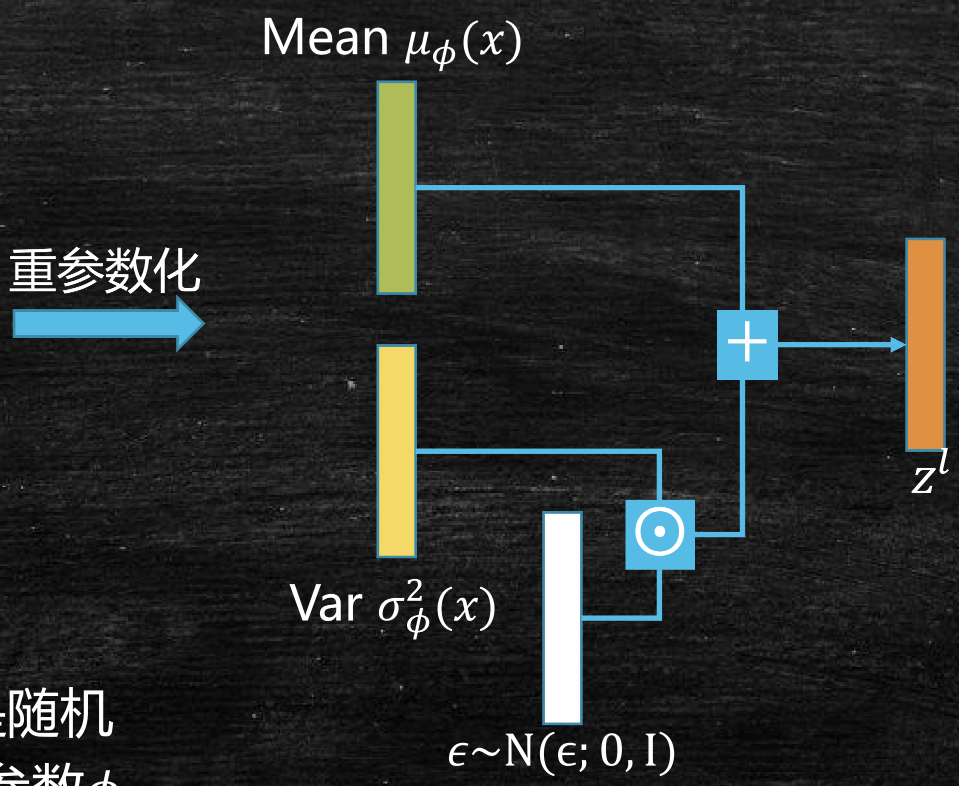
最小化Prior matching term： $q_\phi(z|x) = N\left(z; \mu_\phi(x), \sigma_\phi^2(x)I\right) \rightarrow p(z) = N(z; 0, I)$

KL散度趋近于0，mean向量和var向量分别趋近于全0向量和全1向量

VAE模型结构

重参数化技巧是什么？有什么好处？





$z\sim N\left(z;\mu_\phi(x),\sigma_\phi^2(x)I\right)$分布采样是随机过程，随机过程中包含要优化的参数$\phi$，但是随机采样过程对参数$\phi$ 不可导

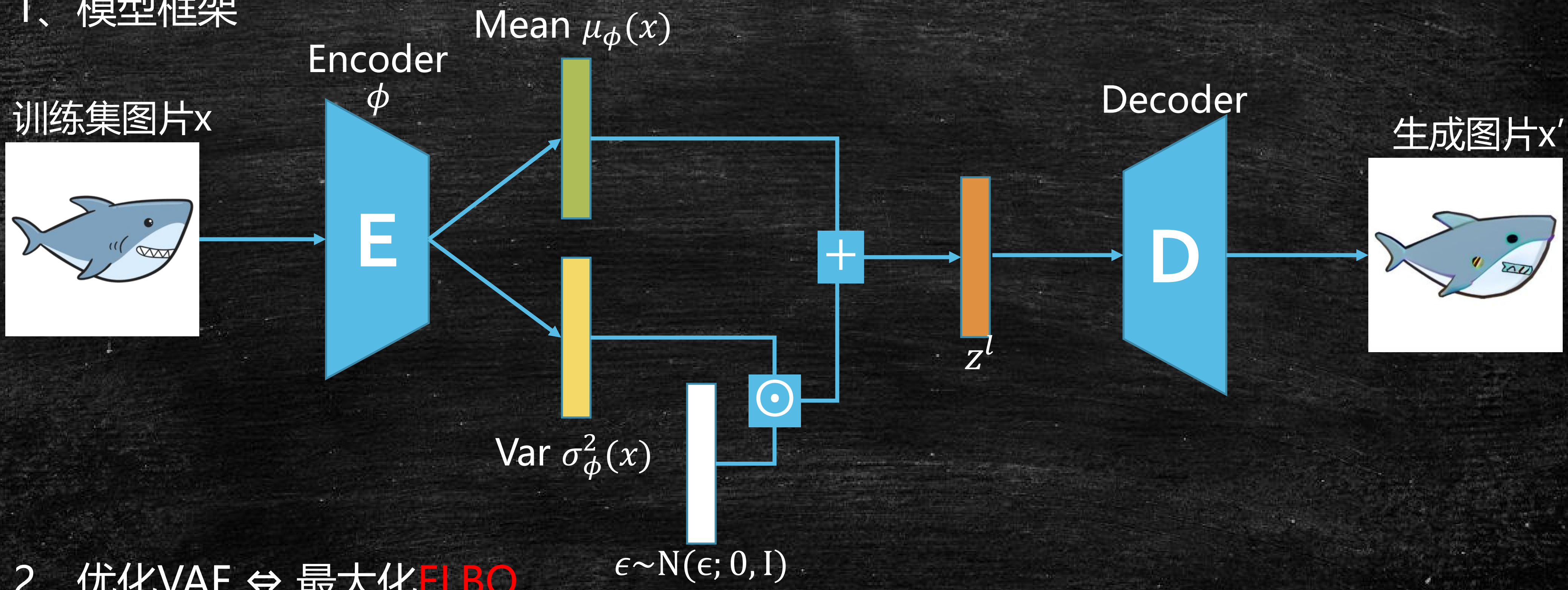
$z^l=\mu_\phi(x)+\sigma_\phi(x)\odot\epsilon$

$\odot$代表元素相乘，这种计算下不改变z的分布

这样参数ϕ 被剥离出了随机采样过程，含参的随机采样变成了不含参采样，这就是重参数化过程，这样参数ϕ 可导

VAE模型小结

1、模型框架



2、优化VAE ⇔ 最大化ELBO

3、ELBO包含两部分，一部分Reconstruction term反映decoder从隐变量重建图片的能力，一部分Prior matching term反映encoder将图片映射到指定隐变量分布的能力

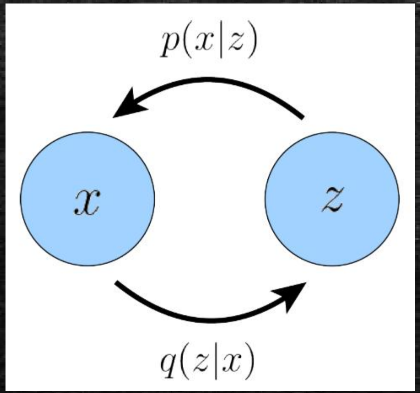
4、重参数化技巧将模型参数剥离出随机采样过程，使其可导

5、生成时，只需要从高斯分布随机采样一个因变量z，经过decoder，就能生成图片啦

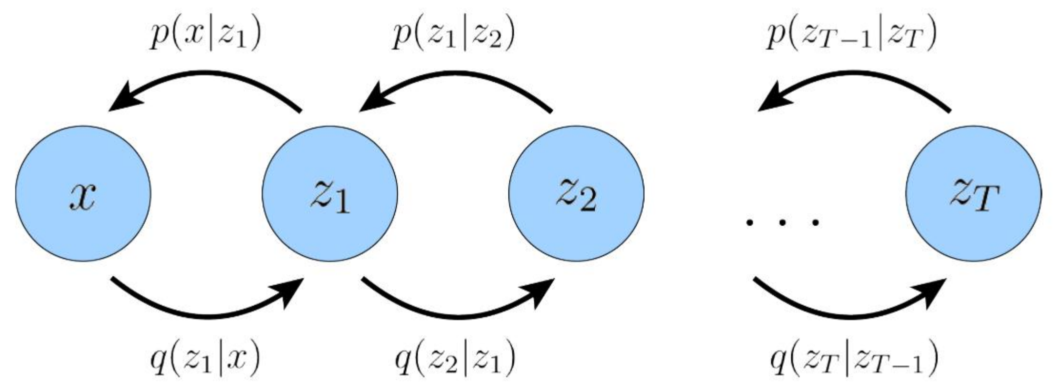
MHVAE理论推导

Markovian Hierarchical Variational AutoEncoder

从VAE到MHVAE



级联
Hierarchical
马尔科夫链
Markovian



$$\begin{aligned}\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} \\ &= \log \int \frac{p(\boldsymbol{x}, \boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} \\ &= \log \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \\ &\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right]\end{aligned}$$

边缘化

×1

期望的定义

琴生不等式

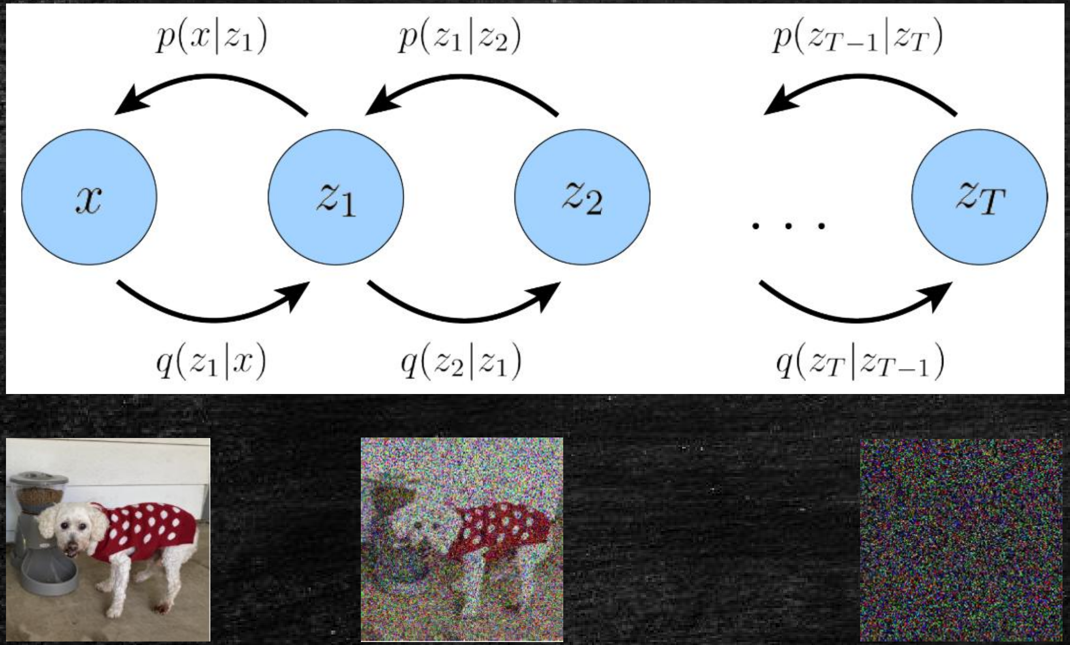
$$\begin{aligned}\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) d\boldsymbol{z}_{1:T} \\ &= \log \int \frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})} d\boldsymbol{z}_{1:T} \\ &= \log \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\right] \\ &\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\right]\end{aligned}$$

VDM理论推导

Variational Diffusion Models

从MHVAE到VDM

MHVAE



+

三个限制：

- ✓ 数据x和所有隐变量z_t的维度相同
- ✓ 所有的encoder $q(z_t|z_{t-1})$都不需要学习，而是预定义好的高斯分布模型，就是z_t状态为以z_{t-1}为均值的高斯分布
- ✓ 最终T状态的分布z_T为标准高斯分布

=

VDM

从MHVAE到VDM

VDM三个限制：

① 数据x和所有隐变量z_t的维度相同

② 所有的encoder $q(z_t|z_{t-1})$都不需要学习，而是预定义好的高斯分布模型，就是z_t状态为以z_{t-1}为均值的高斯分布

③ 最终T状态的分布z_T为标准高斯分布

符号表示替换：

$\mathrm{x} \rightarrow x_0$

$z_t \rightarrow x_t$

限制①②：

人为设定$q(x_t|x_{t-1})$满足如下高斯分布

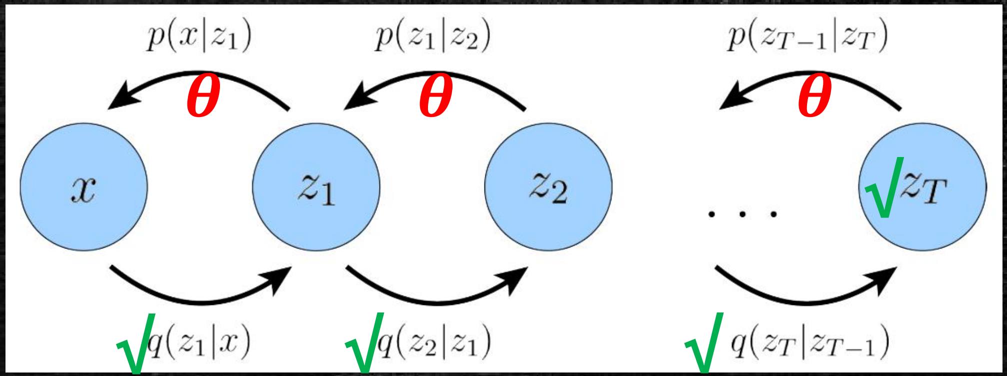
$$q(x_t|x_{t-1})\sim N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$$

α 是超参，一般人为指定，例如SD中的noise schedule，也可以通过模型学习

限制③：

$$p(x_T)\sim N(x_T; \mathbf{0}, I)$$

x_0 … x_t … x_T



VDM数学推导

VDM的ELBO证明
和MHVAE的证明只是把符号表示替换了一下

$$\boldsymbol{x} \to \boldsymbol{x}_0$$
$$\boldsymbol{z}_t \to \boldsymbol{x}_t$$

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) d\boldsymbol{z}_{1:T} \\
&= \log \int \frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})} d\boldsymbol{z}_{1:T} \\
&= \log \mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T})}{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\right] \\
&\geq \mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T})}{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\right]
\end{aligned}$$

边缘化

×1

期望的定义

琴生不等式

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \\
&= \log \int \frac{p(\boldsymbol{x}_{0:T}) q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \\
&= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \\
&\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]
\end{aligned}$$

VDM数学推导

VDM的ELBO拆解，看其具体的含义(页1/2)

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

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

$$\begin{aligned}
p(x_{0:T}) &= p(x_T)p(x_{0:T-1}|x_T) && \text{链式法则}\\
&= p(x_T)p(x_{T-1}|x_T)p(x_{0:T-2}|x_{T-1},x_T)\\
&= p(x_T)p(x_{T-1}|x_T)p(x_{0:T-2}|x_{T-1}) && \text{马尔科夫性质}\\
&\vdots\\
&= p(x_T)p(x_{T-1}|x_T)\ldots p(x_0|x_1)\\
&= \boxed{p(x_T)\Pi_{t=1}^{T}p(x_{t-1}|x_t)}
\end{aligned}$$

$$\begin{aligned}
q(x_{1:T}|x_0) &= q(x_{2:T}|x_1)q(x_1|x_0) && \text{链式法则}\\
&= q(x_{3:T}|x_2,x_1)q(x_2|x_1)q(x_1|x_0)\\
&= q(x_{3:T}|x_2)q(x_2|x_1)q(x_1|x_0) && \text{马尔科夫性质}\\
&\vdots\\
&= q(x_T|x_{T-1})\ldots q(x_2|x_1)q(x_1|x_0)\\
&= \boxed{\Pi_{t=1}^{T}q(x_t|x_{t-1})}
\end{aligned}$$

返回P25

VDM数学推导

VDM的ELBO拆解，看其具体的含义(页1/2)

$$
\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] \\
&= \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]
\end{aligned}
$$

条件概率+马尔科夫性质

分子分母连乘里各拆出一项

连乘部分的符号都减1

和的期望=期望的和

VDM数学推导

VDM的ELBO拆解，看其具体的含义(页2/2)

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

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

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

和的期望=期望的和

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

删掉无关的变量

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

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

见下两页

VDM数学推导

VDM的ELBO拆解，Prior matching term最后一步推导：

$$\mathrm{E}_{q(x_{T-1},x_T|x_0)}\left[log\frac{p(x_T)}{q(x_T|x_{T-1})}\right]$$

Prior matching term

$$=\iint[log\frac{p(x_T)}{q(x_T|x_{T-1})}]q(x_{T-1},x_T|x_0)dx_{T-1}dx_T$$

➡ 多元函数的期望定义

$$=\iint[log\frac{p(x_T)}{q(x_T|x_{T-1})}]q(x_{T-1}|x_0)dx_{T-1}q(x_T|x_{T-1},x_0)dx_T$$

➡ 链式法则

$$=\int[\mathrm{E}_{q(x_T|x_{T-1})}log\frac{p(x_T)}{q(x_T|x_{T-1})}]q(x_{T-1}|x_0)dx_{T-1}$$

➡ 期望的定义+马尔科夫性质

$$=-\int[D_{KL}(q(x_T|x_{T-1})||p(x_T)]q(x_{T-1}|x_0)dx_{T-1}$$

➡ KL散度的定义

$$=-\mathrm{E}_{q(x_{T-1}|x_0)}[D_{KL}(q(x_T|x_{T-1})||p(x_T)]$$

➡ 期望的定义

VDM数学推导

VDM的ELBO拆解，Consistency term最后一步推导：

$$\mathrm{E}_{q(x_{t-1},x_t,x_{t+1}|x_0)}\left[log\frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}\right]$$

Consistency term

$$=\iiint [log\frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}]q(x_{t-1},x_t,x_{t+1}|x_0)dx_{t-1}dx_tdx_{t+1}$$

→ 多元函数的期望定义

$$=\iiint [log\frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}]q(x_{t-1},x_{t+1}|x_0)dx_{t-1}dx_{t+1}q(x_t\,|x_{t-1},x_{t+1},x_0)dx_t$$

→ 链式法则

$$=\iint [\mathrm{E}_{q(x_t|x_{t-1})}log\frac{p_\theta(x_t|x_{t+1})}{q(x_t|x_{t-1})}]q(x_{t-1},x_{t+1}|x_0)dx_{t-1}dx_{t+1}$$

→ 期望的定义+马尔科夫性质

$$=-\iint [D_{KL}(q(x_t|x_{t-1})||p_\theta(x_t|x_{t+1})]q(x_{t-1},x_{t+1}|x_0)dx_{t-1}dx_{t+1}$$

→ KL散度的定义

$$=-\mathrm{E}_{q(x_{t-1},x_{t+1}|x_0)}[D_{KL}(q(x_t|x_{t-1})||p_\theta(x_t|x_{t+1})]$$

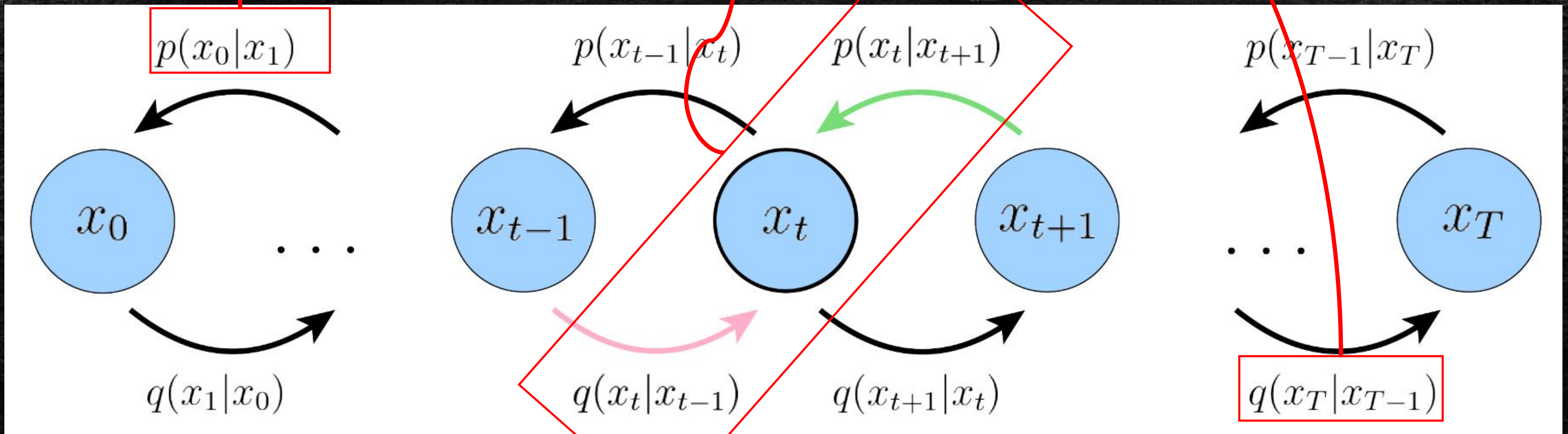
→ 期望的定义

VDM数学推导

VDM的ELBO拆解，看其具体的含义

$$
\begin{aligned}
\log p(\boldsymbol{x}) &\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T))\right]}_{\text{prior matching term}} \\
&\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}))\right]}_{\text{consistency term}}
\end{aligned}
$$

对比VAE：
都有Reconstruction term和
prior matching term
新增consistency term，且该项占主导，因为包含很多子项



VDM数学推导

VDM的ELBO拆解，看其具体的含义

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

包含x_{t-1}和x_{t+1}两个随机变量去预估x_t
用蒙特卡洛方法计算这一项期望误差，两个随机变量估计会方差较大

有办法把2个随机变量变成1个吗？
可以！

通过如下的贝叶斯公式：

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

VDM数学推导

2个随机变量变成1个随机变量推导(1/2页)：

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

→ ELBO

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

→ 链式法则+马尔科夫性质　　之前推导过

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

→ 分子分母连乘里各拆出一项

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

→ 马尔科夫性质

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

→ $\log(AB) = \log(A) + \log(B)$

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

→ 贝叶斯公式

VDM数学推导

2个随机变量变成1个随机变量推导(2/2页)：

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

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

➡ $\log(AB) = \log(A) + \log(B)$

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

➡ $\log(A) + \log(B) = \log(AB)$

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

➡ $E[log(AB)] = E[log(A)]+E[log(B)]$

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

➡ 删掉无关的变量

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

➡ KL散度的定义

VDM数学推导

VDM的ELBO拆解，Denoising matching term最后一步推导：

$$\mathrm{E}_{q(x_t, x_{t-1}|x_0)}\left[log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\right]$$

Denoising matching term

推导和Prior matching term最后一步推导类似

$$= \iint \left[log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\right] q(x_t, x_{t-1}|x_0)dx_{t-1}dx_t$$

多元函数的期望定义

$$= \iint \left[log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}\right] q(x_{t-1}\,|x_t, x_0)dx_{t-1}q(x_t\,|x_0)dx_t$$

链式法则

$$= \int [\mathrm{E}_{q(x_{t-1}|x_t, x_0)}log\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)}]q(x_t\,|x_0)dx_t$$

期望的定义

$$=-\int [D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))]q(x_t\,|x_0)dx_t$$

KL散度的定义

$$=-\mathrm{E}_{q(x_t\,|x_0)}[D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))]$$

期望的定义

VDM数学推导

VDM的ELBO拆解， 2个随机变量变成1个随机变量

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

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

对比之前2个随机变量公式：

- Reconstruction term和prior matching term保持一致
- 原来的consistency term，变成了denoising matching term
- 原来consistency term由x_{t-1}和x_{t+1}2个随机变量去预估x_t
- 现在denoising matching term仅由x_{t+1}1个随机变量去预估x_t，采用蒙特卡洛方法时，估计方差变小

说明：
以上公式仅仅用到了马尔科夫定理，所以适用于所有MHVAE模型，而不仅仅局限于VDM模型。

VDM数学推导

VDM的ELBO拆解， 2个随机变量变成1个随机变量

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

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

对比VAE推导的ELBO拆解公式：

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] = \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))}_{\text{prior matching term}}$$

结论：
当T=1时，denosing matching term=0， x_0用x表示， x_1用z表示，VDM的ELBO拆解与VAE的ELBO拆解结果一致。

VDM数学推导

VDM的ELBO拆解， 1个随机变量版本

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

无可优化参数

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

denoising matching term由于是多项的求和，所以会在优化目标中占主要部分。所以重点看看这部分吧！

$p_\theta(x_{t-1}|x_t)$这个分布是模型要去学习的。
$q(x_{t-1}|x_t, x_0)$这个分布就是我要让模型学习的目标，尽可能与这个分布相等。换句话说：$q(x_{t-1}|x_t, x_0)$就是模型要学习的ground-truth。

$q(x_{t-1}|x_t, x_0)$这个GT怎么算呢？

VDM数学推导

$q(x_{t-1}|x_t,x_0)$这个ground-truth怎么算呢?

Recall：MHVAE->VDM的三个限制条件之二

所有的encoder $q(x_t|x_{t-1})$都不需要学习，而是预定义好的高斯分布模型，就是x_t为以x_{t-1}为均值的高斯分布，用数学表达就是：

$$q(x_t|x_{t-1}) \sim N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$$

看到$q(x_t|x_{t-1})$，求$q(x_{t-1}|x_t,x_0)$，贝叶斯公式!

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

另外几项：$q(x_t|x_{t-1},x_0)$，$q(x_{t-1}|x_0)$和$q(x_t|x_0)$具体怎么计算呢?

再Recall：重参数化技巧

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \boldsymbol{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{I})$$

VDM数学推导

$q(x_{t-1}|x_t, x_0)$这个ground-truth怎么算呢?

用递推算$q(x_t|x_0)$:

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

重参数化

x_{t-1}重参数化+迭代

系数合并

高斯分布：和的方差等于方差的和
和的均值等于均值的和

系数合并

$x_{t-2}, \ldots, x_1$重参数化+多次迭代直到用x_0来表示

参数表示替换：$\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$

$q(x_t|x_0)$为高斯分布

VDM数学推导

$q(x_{t-1}|x_t,x_0)$这个ground-truth怎么算呢？(推导1/2页)

用同样的方法也能得到：$q(x_{t-1}|x_0) \sim N(x_{t-1};\sqrt{\bar{\alpha}_{t-1}}x_0,(1-\bar{\alpha}_{t-1})I)$
$q(x_t|x_{t-1},x_0)$，$q(x_{t-1}|x_0)$和$q(x_t|x_0)$三项带入贝叶斯公式：

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

→ 贝叶斯公式

$$= \frac{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\alpha_t}\boldsymbol{x}_{t-1},(1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1};\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0,(1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,(1-\bar{\alpha}_t)\mathbf{I})}$$

→ 带入三个高斯分布

$$\propto \exp\left\{-\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{2(1-\alpha_t)}+\frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_{t-1})}-\frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_t)}\right]\right\}$$

→ 带入高斯分布的表达式，隐去系数

$$= \exp\left\{-\frac{1}{2}\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{1-\alpha_t}+\frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_{t-1}}-\frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_t}\right]\right\}$$

→ 提出通项1/2

$$= \exp\left\{-\frac{1}{2}\left[\frac{(-2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}+\alpha_t\boldsymbol{x}_{t-1}^2)}{1-\alpha_t}+\frac{(\boldsymbol{x}_{t-1}^2-2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0)}{1-\bar{\alpha}_{t-1}}+C(\boldsymbol{x}_t,\boldsymbol{x}_0)\right]\right\}$$

→ 平方开方

$$\propto \exp\left\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}}{1-\alpha_t}+\frac{\alpha_t\boldsymbol{x}_{t-1}^2}{1-\alpha_t}+\frac{\boldsymbol{x}_{t-1}^2}{1-\bar{\alpha}_{t-1}}-\frac{2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right]\right\}$$

→ 先忽略常数

$$= \exp\left\{-\frac{1}{2}\left[\left(\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\}$$

→ 合并同类项系数

VDM数学推导

$q(x_{t-1}|x_t, x_0)$这个ground-truth怎么算呢？(推导2/2页)

$$= \exp\left\{-\frac{1}{2}\left[\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\right\}$$

→ 合并同类项系数

$$= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)}{\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}}\boldsymbol{x}_{t-1}\right]\right\}$$

→ 提取系数

$$= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\}$$

→ 系数化简

$$= \exp\left\{-\frac{1}{2}\left(\frac{1}{\boxed{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\boxed{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}\boldsymbol{x}_{t-1}\right]\right\}$$

→ 凑成高斯分布表达式的系数形式

$$\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t,\boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})$$

→ 把之前忽略的常数拿回来，正好能配成一个高斯分布的表达式，得到分布的均值和方差

高斯分布：$\mathrm{x} \sim N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)[x^2 - 2\mu x + \mu^2]) \propto \exp(-\frac{1}{2}\left(\frac{1}{\sigma^2}\right)[x^2 - 2\mu x])$

VDM数学推导

$q(x_{t-1}|x_t, x_0)$这个ground-truth的特性?

$$\mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})$$

- ◻ 高斯分布
- ◻ 方差Σ_q(t)只跟的α_t有关，每一步的α_t已知
- ◻ 均值$\mu_q(x_t, x_0)$是关于x_t, x_0的函数

Recall：求这个是干啥啊？因为如下ELBO拆解的第三项需要用模型拟合$q(x_{t-1}|x_t, x_0)$

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

模型优化目标：最小化$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t, x_0)$的KL散度

VDM数学推导

最小化$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t,x_0)$的KL散度

$$\mathcal{N}(\boldsymbol{x}_{t-1};\underbrace{\frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})\boldsymbol{x}_t+\sqrt{\bar\alpha_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar\alpha_t}}_{\mu_q(\boldsymbol{x}_t,\boldsymbol{x}_0)},\underbrace{\frac{(1-\alpha_t)(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})$$

- 高斯分布
- 方差$\Sigma_q(\mathrm{t})$只跟的α_t有关， 每一步的α_t已知
- 均值$\mu_q(x_t,x_0)$是关于x_t,x_0的函数

模型优化目标：最小化$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t,x_0)$的KL散度

根据ground-truth的前两条特性，我们也可以给我们要预估的分布以下两条性质：

- ✓ 定义$p_\theta(x_{t-1}|x_t)$为高斯分布
- ✓ 定义方差与$q(x_{t-1}|x_t,x_0)$的相等， 等于$\Sigma_q(\mathrm{t})$

VDM数学推导

最小化$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t,x_0)$的KL散度
下标θ和q分别代表模型预估的分布和ground-truth分布

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

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

方差相等，均为$\Sigma_q(t)$

$$=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_q(t)|}{|\boldsymbol{\Sigma}_q(t)|}-d+\operatorname{tr}(\boldsymbol{\Sigma}_q(t)^{-1}\boldsymbol{\Sigma}_q(t))+(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right]$$

两个高斯分布的KL散度计算公式

$$=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2}\left[\log 1-d+d+(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right]$$

$tr(I)=d$, d是方差矩阵的维度

$$=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right]$$

简单化简

$$=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2}\left[(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)^T\left(\sigma_q^2(t)\mathbf{I}\right)^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q)\right]$$

$\Sigma_q(t)=\sigma_t^2 I$已知

$$=\underset{\boldsymbol{\theta}}{\arg\min}\,\frac{1}{2\sigma_q^2(t)}\left[\|\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q\|_2^2\right]$$

简单化简

VDM数学推导

最小化$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t,x_0)$的KL散度

$$\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q\right\|_2^2\right]$$

最小化KL散度 ⇔ 拟合$q(x_{t-1}|x_t,x_0)$均值

要拟合的均值$\mu_q(x_t,x_0)$ 如下：

$$\mu_q(x_t,x_0)=\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}$$

$\mu_\theta(x_t,t)$也是关于x_t的表达式，为了接近$\mu_q(x_t,x_0)$的形式，我们将$\mu_\theta(x_t,t)$写成如下公式：

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

这样：最小化KL散度 ⇔ 拟合$q(x_{t-1}|x_t,x_0)$均值 ⇔ 根据x_t和步数t拟合x_0

VDM数学推导

最小化$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t, x_0)$的KL散度

最小化KL散度 ⇔ 拟合$q(x_{t-1}|x_t, x_0)$均值 ⇔ 根据x_t和步数t拟合x_0

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

全部都是简单化简

VDM数学推导

最小化$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t,x_0)$的KL散度

（预测噪声的损失函数表达）

$$\arg\min_{\theta} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) = \arg\min_{\theta} \frac{1}{2\sigma_q^2(t)}\left[\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q\right\|_2^2\right]$$

Recall：

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

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

$$\mu_q(x_t,x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}$$

$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\frac{x_t-\sqrt{1-\bar{\alpha}_t}\epsilon_0}{\sqrt{\bar{\alpha}_t}}}{1-\bar{\alpha}_t}$$

简单带入

$$= \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\epsilon_0$$

合并同类项 & $\bar{\alpha}_t = \bar{\alpha}_{t-1}\alpha_t$

同理：我们把模型预测的$\mu_\theta(x_t,t)$也写成上面的形式：

$$\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boxed{\hat{\epsilon}_\theta(x_t,t)}$$

VDM数学推导

最小化$p_\theta(x_{t-1}|x_t)$和$q(x_{t-1}|x_t,x_0)$的KL散度（关于预测噪声的损失函数表达）

$$\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)}\left[\left\|\boldsymbol{\mu}_{\boldsymbol{\theta}}-\boldsymbol{\mu}_q\right\|_2^2\right]$$

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

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

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

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

全部都是简单化简

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

这样：最小化KL散度 ⇔ 拟合$q(x_{t-1}|x_t,x_0)$均值 ⇔ 拟合初始图像x_0 ⇔ 拟合噪声ϵ_0

VDM数学推导

Inference阶段:

每次来一个x_t和t，模型预测得到加的噪声：$\hat{\epsilon}_\theta(x_t, t)$

然后就可以算x_{t-1}的均值，然后标准高斯分布上随机采样出一个噪声$\mathbf{z}$，方差$\Sigma_q(\mathrm{t}) = \sigma_{\mathrm{t}}{}^2 I$已知，用重参数化技巧得到$x_{t-1}$：

$$x_{t-1}(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_{\mathrm{t}}}\sqrt{\alpha_t}} \hat{\epsilon}_\theta(x_t, t) + \sigma_{\mathrm{t}}\mathbf{z}$$

经过t次迭代，就能得到没有噪声的x_0

DDPM训练和采样流程:

Algorithm 1 Training

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

Algorithm 2 Sampling

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

小结

1、VDM = 马尔可夫级联VAE（MHVAE） + 3个限制条件

2、优化VDM和优化VAE一样，都是最大化ELBO

3、拆解VDM的ELBO能得到Reconstruction term、prior matching term和denoising matching term三项，第三项为VDM相比VAE新增，该项因为是多项求和，占据损失函数的主导。当T=1时，VDM和VAE的ELBO相等。

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$$

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

4、 $q(x_{t-1}|x_t, x_0)$就是模型要学习的ground-truth。这一项可以计算得到。也是高斯分布，满足：

$$\mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})$$

小结

5、模型优化目标：最小化KL散度 ⇔ 拟合$q(x_{t-1}|x_t, x_0)$均值 ⇔ 拟合初始图像x_0 ⇔ 拟合噪声ϵ_0

$$\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_q^2(t)} \left[\|\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\mu}_q\|_2^2\right]$$

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

6、Inference时，根据模型预测计算t-1步均值，再在标准高斯分布上随机采样出一个噪声$\mathbf{z}$，方差$\Sigma_q(\mathrm{t}) = \sigma_{\mathrm{t}}{}^2 I$已知，用重参数化技巧得到$x_{t-1}$：

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_{\mathrm{t}}}\sqrt{\alpha_t}} \hat{\epsilon}_\theta(x_t, t)$$

$$x_{t-1}(x_t, t) = \mu_\theta(x_t, t) + \sigma_{\mathrm{t}}\mathbf{z} = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_{\mathrm{t}}}\sqrt{\alpha_t}} \hat{\epsilon}_\theta(x_t, t) + \sigma_{\mathrm{t}}\mathbf{z}$$

结语

基础理论大厦已经建成，但是仅凭如此还不够！

- 迭代步数太多，生成速度太慢（DDIM）
- 隐空间维度低，无法生成高清大图（SD使用的LDM）
- 加噪过程能否优化，固定方差，只预测均值真的是最优解吗（IDDPM）
- 如何控制生成的方向，得到想要的生成内容（CG、CFG）
- 什么网络结构能得到更好的生成效果（U-net、DiTs）
- 更多实验得到的工程优化（DDPM的L-simple）
- ……